

Budget let-downs

With the United States facing tough economic times, science gets small change and less sense of direction from President Bush's research and development budget proposal for 2005.

The generally lacklustre trajectory of the US research budget announced last week by President Bush hardly comes as a surprise, given the country's mountainous economic woes. The 2005 budget allows no growth, after inflation is taken into account, at the National Institutes of Health (NIH) or the National Science Foundation (NSF). Most other agencies do even less well, and those with any connection to environmental science, such as the Environmental Protection Agency and the National Oceanic and Atmospheric Administration, fare even worse.

Despite the protestations of biomedical-research lobbyists, the NIH was expected to enter a period of consolidation after five years of rapid growth. Congress may now look for direct results from an agency that now spends \$28 billion a year. But the treatment of the NSF, which funds most non-biomedical research at US universities, is disappointing and inconsistent. In late 2002, Bush signed an authorization bill that allowed for the doubling of the NSF's funding over five years. It was wrong for him to sign this bill, which was about long-term planning, and then ignore its contents.

The reorientation of NASA towards a revived human-spaceflight programme had already produced some scientific casualties ahead of the budget's release — most notably, a shortened life for the Hubble space telescope. But Hubble is only one of a number of odd choices

made to redirect support towards a possible manned Moon mission. These should now be scrutinized in Congress.

Elsewhere, economic policy is coming home to roost. Despite a boost to combat mad cow disease, resources will remain short at the US Department of Agriculture and at the Centers for Disease Control and Prevention, which deals with global infectious diseases. With deficits as far as the eye can see, pressing scientific challenges facing the government evidently cannot attract sufficient resources.

The budget also shows few signs of effective interagency cooperation. The White House Office of Science and Technology Policy is supposed to orchestrate this, but its footprints on the budget are hard to see. To give just one example: NASA has set aside no funds for a space-based physics experiment that it planned in collaboration with the Department of Energy, the main US physics agency. But the energy department has budgeted to continue with it.

There is little indication in the research budget of direction from the centre. Previous presidents sought to coordinate action in areas such as climate change, nanotechnology and mathematics. Scientists don't always welcome these special initiatives, but in these areas they have been effective at getting things moving on US university campuses. No sign of that this year. If this turns out to be George W. Bush's last budget, he will have gone out with a whimper. ■

Keys to capacity

The task of advancing science in developing countries is beyond any one nation or organization. How can scientists help?

Two reports just published have reached the same conclusion: developing countries need to grow their science and technology bases if they are to exploit new knowledge and technologies. This is not a luxury but an absolute necessity; otherwise the technology gap between them and industrialized countries will widen, and they will be impotent to roll back the plagues of killer diseases, or to generate sustainable agriculture, energy and water.

Inventing a Better Future (<http://www.interacademycouncil.net>) is by the InterAcademy Council, a body set up four years ago by 90 national science academies to advise the United Nations and other international bodies (see page 577). The second report, *Science, Technology and Innovation* (http://www.cid.harvard.edu/cidtech/interim_report.doc) is by the science task force of the United Nations Millennium Project, a plan adopted by world leaders in 2000 that sets hopelessly unrealistic goals for slashing the burden of poverty, hunger and disease by 2015, but has at least put them high on the agenda.

The reports' remedies largely rely on collective action by individual governments and the international community. This is essential, but bureaucracy and inertia will mean that we shouldn't expect progress anytime soon. Meanwhile, the international scientific community could do more to help expedite the same goals.

The Bill and Melinda Gates Foundation, the UK Wellcome Trust and other charities already generously support centres in the developing world. And the Howard Hughes Medical Institute, the

National Institutes of Health's Fogarty International Center, and many research agencies worldwide also operate schemes to support talented researchers and their institutions.

Yet good databases of the support available for capacity building and of promising talent do not exist. They would surely reveal opportunities for agencies and Northern labs to work better together to concentrate support on priority areas, and on the best science, to nurture enduring centres of excellence.

Furthermore, there is a case for one or more lightweight, international coordinating initiatives, including research organizations, donors and the private sector. Such initiatives could also set standards for peer review, and for ensuring that research funds don't end up in the hands of corrupt university or government officials — a problem rife in Africa. Agreements on competitive salaries are also urgently needed to stem the brain drain.

One immediate impact could be achieved cheaply by providing the best research centres in developing countries with high-speed satellite Internet access. Scientists in the few African laboratories that have it report not just greater access to information and better networking with colleagues, but vastly greater success in applying for overseas grants because they are more up to speed.

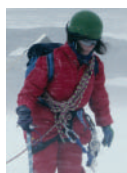
Putting science in developing countries on a stronger and more transparent footing would attract investment and raise the bar for the foreign ministries of Western governments who disburse large sums, but often on the basis of anything but good science. ■





Cattle alert

International panel calls for greater BSE surveillance in US
p575



Natural treasures

Bioprospecting agreements run into trouble
p576



Fishy peril

Virus threatens carp stocks worldwide
p577



Old bones

Appeals court decides on Kennewick Man
p578

Virologists call for vaccination in bid to beat bird flu epidemic

Alison Abbott

Bird vaccination is needed to bring south-east Asia's epidemic of chicken flu under control, experts have concluded.

On 5 February the United Nations Food and Agriculture Organization (FAO) advised governments in affected areas that mass culling of birds (see *Nature* 427, 472–473; 2004) is failing to halt the disease and that vaccination of targeted poultry flocks is required as well.

The advice followed an emergency meeting on vaccination strategies held by the FAO and other international bodies in Rome on 3–4 February. FAO officials are now working out the cost of vaccination, which would involve trained personnel administering what could be hundreds of millions of shots.

They will present a package of control plans, which include different levels of vaccination, at a regional FAO meeting in Bangkok later this month. The cost will vary enormously between commercial farms and remote villages, says Joseph Domenech, head of the FAO's animal health service — and it is unclear who will pay.

Chicken flu was relatively rare until 1997. But in the past six years more than 100 million birds have been infected by the disease or culled in its wake — half of them in southeast Asia this year — as farming becomes more intensive and reliant on the frequent movement of stock.

Until now, agricultural authorities have favoured culling over vaccination, in part because vaccinated animals can still harbour and shed infectious virus. Until recently it had not been possible to differentiate between infected and uninfected vaccinated animals, so vaccinated animals were still subject to trade bans.

But vaccination could help bring the latest outbreak under control by reducing the viral load in the environment, says Juan Lubroth, an avian-influenza expert at the FAO.

"Vaccination reduces the level of excretion of virus in infected animals, and also raises the threshold of viral load required for infectivity," explains Ilaria Capua, director of virology at the Veterinary Public Health Institute in Padua, Italy, whose vaccination



Chickens may be vaccinated in an effort to slow southeast Asia's avian flu epidemic.

strategy has already controlled two outbreaks of chicken flu in Italy.

Capua's strategy, known as DIVA (differentiating infected from vaccinated animals) allows chickens that are vaccinated, but still carrying the virus, to be identified by a molecular diagnostic test.

Avian influenza viruses have two key surface proteins. One is the haemagglutinin antigen, which has 15 subtypes, H5 and H7 being most frequently associated with pathogenicity. The other is neuraminidase (N), which has nine subtypes and is less important in protection.

The vaccine used in a DIVA strategy is made from strains with the same H subtype as the circulating 'field' strain, but with a different N subtype. In the 2000–01 Italian outbreak of an H7N1 bird flu, a vaccine against an H7N3 strain was used. Vaccinated birds developed antibodies to the H7 and N3 proteins; birds infected with the circulating H7N1 strain also made antibodies to N1. A test to detect the anti-N1 antibodies was developed by scientists at Capua's institute, and the DIVA strategy was validated and approved by the Paris-based World Organisation for Animal Health and the European Union.

Confidence in the test, the operation of which was regularly checked by European

Commission officials, quickly led to the lifting of a trade ban on products from vaccinated, uninfected animals.

"We will consider using different vaccination strategies for different situations," says Lubroth. In some circumstances, the FAO will suggest using 'sentinel' birds — unvaccinated birds kept within the vaccinated flocks — to monitor infection simply by the display of clinical symptoms. This is difficult to check, however, as farmers trying to hide infection could swap sick sentinel birds for healthy ones.

The bird vaccines cost just US\$2.5–5 cents per injected shot but even so a vaccination programme will be expensive as the number of vaccinations could run into billions. However, according to Capua, it could still be the best option if the flu continues to spread.

It would be hard to introduce the DIVA strategy in southeast Asia, however, as some affected countries have few facilities for testing for antibodies.

Capua is worried that some of the poorer countries in southeast Asia may not be able to monitor vaccinated birds carefully enough. "We have good vaccination tools to help the eradication effort, but if we don't use them properly it could make the situation worse," she warns.

XINHUA CHEN SHUGEN/AP

Retraction signals end of cell-biology debate

Erika Check, Washington

A biology paper that could have overturned a widely accepted theory on cell signalling has been retracted 15 months after it was published.

The retraction has rocked the cell-biology community and, say observers, has effectively ended the career of Siu-Kwong Chan, one of the paper's co-authors.

Gary Struhl, a Howard Hughes Medical Institute (HHMI) Investigator based at Columbia University, New York, and the senior author on the paper, issued the retraction on 6 February (*G. Struhl Cell* 116, 481; 2004). In it, Struhl notes that Chan, who was a postdoc in his lab, has admitted misreporting or failing to perform crucial experiments described in the original paper (S.-K. Chan and G. Struhl *Cell* 111, 265–280; 2002).

Struhl discovered a problem when he repeated some of Chan's experiments. When he didn't get the same results as Chan, Struhl says that he confronted his former postdoc, who had by this time moved to the Albert Einstein College of Medicine in the Bronx.

"When confronted with this discrepancy, S.-K. Chan informed me that most of the results shown in Figures 2D, 4, and 5, including the negative control shown in Figure 5B, were either not performed or gave different results than presented in the paper," Struhl wrote in the retraction. "I therefore withdraw this paper and the conclusions it reports." He declined to comment further.

Struhl and Chan were working on a family of proteins known as Wnt, which plays a crucial part in a signalling pathway involved in cellular development. Chan said that he had



It could have changed our view of some cancers — but this *Cell* paper was retracted last week.

found evidence that contradicted conventional wisdom about how the Wnt pathway works, and showed the data to Struhl. The pair then worked for the next five years on different components of the project, before publishing their results in October 2002.

The paper caused a stir when it first appeared, as many prominent cell biologists — including Harold Varmus, president of the Memorial Sloan-Kettering Cancer Center in New York — didn't accept its hypothesis.

The work focused on two key proteins called Armadillo and TCF in fruitflies. Conventionally, Armadillo was thought to work

by combining with TCF inside the cell nucleus. But Struhl and Chan's paper suggested that Armadillo either shuttled TCF out of the nucleus, or activated TCF outside the nucleus. The details of this process have major implications, because malfunction of the Wnt pathway causes many human cancers. "It's crucial to know how this pathway works before we can figure out how to interfere with it," explains Varmus.

Thanks to the retraction and the findings of other researchers, most cell biologists now accept the original explanation of how the Wnt pathway works.

The HHMI, which has funded Struhl's work since 1986, as well as supporting Chan's work on the 2002 paper in *Cell*, says that it is now investigating the episode in cooperation with Columbia. Officials at the university did not respond to requests for comment.

And scientists who previously worked with Chan have had to ask him whether he deceived them as well. In the mid-1990s, Chan co-authored several papers with Richard Mann, his graduate thesis adviser, who also works at Columbia. Mann says he is confident that Chan's graduate work was solid, because it has been built on and confirmed by other scientists in his lab and elsewhere. Mann adds that Chan worked on those projects with other scientists who said the work was performed honestly.

According to Karen Gardner, a spokeswoman for the Albert Einstein College of Medicine, Chan resigned his post on 21 January, less than three months after he started there. Chan did not reply to an e-mail seeking comment. ■

NASA steels itself for rough ride over Hubble's demise

Tony Reichhardt, Washington

Mounting criticism of NASA's decision to cancel an upgrade of the Hubble Space Telescope will surface at a congressional hearing this week. Critics are expected to charge that the agency has not been honest about its reasons for the cancellation.

NASA administrator Sean O'Keefe has repeatedly said that astronaut safety — not money — is behind the decision. The 2006 space shuttle mission would have installed two new science instruments and replaced gyroscopes and batteries in the 14-year-old telescope to extend its life to 2010. Without this repair, one of NASA's most productive scientific missions could fail as early as 2007.

Many astronomers think that the real reason for cancelling the mission is to conserve resources to fund President George Bush's proposal to return astronauts to the

Moon. The House Committee on Science, chaired by Sherwood Boehlert (Republican, New York), will hold a hearing on 12 February on the plan, with O'Keefe and White House science adviser John Marburger as the main witnesses.

Committee members are expected to raise two reports by an unnamed NASA engineer that are circulating on the Internet. These contend that O'Keefe's statement that a Hubble flight would be riskier than space station flights "cannot be supported" on technical grounds. Hubble's higher orbit actually poses less of a threat of orbital debris hitting the shuttle's fragile insulating tiles, say the reports. And NASA's plan to have another shuttle ready to rescue astronauts in case their vehicle fails could just as well be used for Hubble missions as for space station missions, the reports argue.

Many in the space community have questioned O'Keefe's decision, none more harshly than Robert Zubrin, president of the Mars Society, a group that advocates manned Mars expeditions. "The grounds given for deserting Hubble are irrational, and constitute a form of moral cowardice that if accepted as the basis of space policy, would absolutely prevent any human missions to the Moon, Mars, or anywhere else," Zubrin said in a statement last week.

The controversy pits scientists against NASA at a time when the agency could do with their support for President Bush's widely criticized Moon initiative. It also calls into question the regime of openness that O'Keefe has said he is implementing. *The New York Times* reported on 7 February that the author of the reports wished to remain anonymous for fear of losing his job. ■



High stakes: the United States has been advised to expand its monitoring of cattle herds for disease.

Food panel calls for beefed up response to mad cow disease

Erika Check, Washington

An international panel of experts is urging the US Department of Agriculture (USDA) to take extra measures to block the spread of mad cow disease. But the department looks unlikely to comply, after a domestic advisory group expressed scepticism about the scientists' suggestions.

Ulrich Kihm, Switzerland's chief veterinary officer and chair of the international panel, says that there are sure to be more diseased cows in the United States than the one found in Washington state in December. "There is no question that there are many more out there," he says.

In a report released on 4 February, Kihm's group said that the United States should expand its cattle surveillance programme and put extra restrictions on the types of animal products that can be fed to livestock and pets.

The panel was invited to review the US response to mad cow disease, or bovine spongiform encephalopathy (BSE), shortly after the Washington case was diagnosed. Agriculture secretary Ann Veneman asked it to assess whether the US investigation conformed to international standards and to recommend further steps that might be taken.

But the USDA's Advisory Committee on Foreign Animal and Poultry Diseases, which received the report at a meeting near Washington, wasn't impressed with the panel's findings. "We've had tens and tens of thousands of animals tested since 1997," said committee member Tobin Armstrong, a rancher in Texas. "If you want to measure the extent of the problem, you have to go on those numbers — and the number is one."

The National Cattlemen's Beef Association issued a statement saying that the four overseas members on the five-strong expert panel had underestimated the measures already in place to block BSE. "Many of the panel's recommendations are based on the European model and overlook scientific evidence that clearly demonstrates the long-standing firewalls in place in our country have been effective," it said.

Since December, the USDA has brought in a series of rules to bolster defences against both BSE and its human equivalent, variant Creutzfeldt-Jakob disease. It has banned the slaughter for beef of animals that are too sick to walk, and farmers now cannot sell parts of older cows' central nervous systems as food.

But Kihm's review panel says that's not enough. It argues that other infected cows were probably turned into cow meal before rules enacted in 1997 prohibited the use of cow parts in cattle feed. It adds that feed bans have been difficult to enforce in Europe.

The panel also questioned the USDA's plan to test just 40,000 high-risk cattle over the next year. It says that officials should consider testing all cows over 30 months old in certain high-risk populations, as well as a sample of other, older animals.

"You'd need to test a very large sample of fallen stock to get a good idea of the real situation," says Marc Savey, research director at the French Food Safety Agency. "It should be more like 800,000 to 3 million animals — 40,000 is ridiculous."

Veneman's advisory committee will now review the international team's report and make its own recommendations. ■

Pacific dolphins make waves for US policy on Mexican tuna

Virginia Gewin, Portland

Conservationists are accusing the US government of ignoring its own scientific advisers over its decision to import tuna fish from Mexico.

In December 2002, the Department of Commerce agreed that tuna from Mexico could be labelled 'dolphin safe' and so be imported and sold in the United States.

But a conservation organization last week revealed a cache of documents that suggest this decision was taken against scientific advice.

Biologists want the imports banned because Mexican fishermen target and chase dolphins in order to catch the underlying tuna. They argue that this is contributing to the decline in dolphin numbers in the Pacific Ocean.

The papers were circulated by the Earth Island Institute, a conservation group in San Francisco and one of several plaintiffs suing the government over the issue, after they were released by a district court judge in San Francisco towards the end of last year. They are a "smoking gun," according to the institute's director, David Phillips.

The documents list talking points for briefings prepared by officials for the commerce secretary, Donald Evans. One states: "We've all seen the science. We know that dolphins aren't recovering." Another says that "a determination of 'no significant adverse impact' is not supported by the science".

The issue was analysed by the Southwest Fisheries Science Center, a San Diego-based research arm of the National Marine Fisheries Service (NMFS). Its report, released in August 2002, said that indirect effects of tuna fishing such as stress could explain declining dolphin populations in the Pacific.

A five-member independent expert panel unanimously agreed with these findings, and the current lawsuit is backed by three members of this panel.

But the commerce department got different advice from the Inter-American Tropical Tuna Commission, which is funded by governments to maintain maximum sustainable tuna catches in the Pacific. It contends that if Mexican fishermen adopted new fishing methods, it would damage marine ecosystems further.

NMFS director William Hogarth maintains that the initial import decision was justified, but neither he nor the chief scientist from the Southwest Center would comment further pending the district court hearing on 29 March. ■

Bioprospectors hunt for fair share of profits

Rex Dalton, San Diego

Representatives of almost 200 nations convene in Malaysia this week to try to agree terms for sharing profits from natural molecules and organisms between indigenous peoples, scientists, governments and drug companies.

But as delegates gather in Kuala Lumpur to update the ten-year-old Convention on Biological Diversity, researchers and pharmaceutical companies worldwide are struggling to make bioprospecting for such compounds work.

Even in the United States, which has a sophisticated legal and governance system, benefit-sharing agreements have been slow to take shape. For instance, five years ago a federal court in Washington DC ordered the US National Park Service to develop a plan for sharing the bounty from valuable compounds discovered in its parks. But the

plan has yet to be published and critics are hitting out at the park service for the delay (see box below).

At this week's meeting, which runs until 20 February, delegates from developing nations — many of them rich in diversity — are expected to seek a new, binding treaty that would set stricter rules on benefit sharing. But Europe, Japan and the United States, which has yet to ratify the original convention, are expected to oppose such a move.

Everyone agrees that the existing convention is not working well. Researchers and pharmaceutical companies say it has created bureaucratic impediments to commercializing discoveries. Poor nations and local populations are demanding assurances that any commercial benefits will be shared.

These benefits can take years to realize, however, so companies are reluctant to commit themselves in advance to costly deals.

Faced with slim pickings and the prospect of complex negotiations over intellectual property rights, major drug companies have scaled back bioprospecting, leaving it to smaller biotechnology companies.

Bioprospecting is thought to hold special potential in tropical regions, home to most of the world's biodiversity. But promising natural compounds could be discovered anywhere. Last month, for example, a report by the United Nations University's Institute of Advanced Studies in Tokyo highlighted the bioprospecting potential of Antarctica.

The report said the continent's arid and salty conditions have led to the evolution of tough organisms that offer opportunities for the development of products such as industrial chemicals, drugs and genetic components. It called for a new regulatory framework between the 54 nations that have a say in governing Antarctica, warning that existing arrangements are threatening funding for bioprospecting expeditions and creating the potential for disputes that may mire discoveries in litigation.

Ironically, many scientists who originally supported the convention have been the ones damaged most by it, says Leonard Hirsch, a senior policy adviser at the Smithsonian Institution in Washington DC and joint leader of the US delegation in Kuala Lumpur. But he is optimistic that despite current disagreements, some course can be worked out to facilitate more bioprospecting.

Privately, one negotiator says that Europe may play a key role in determining whether talks begin on a new international treaty covering access and benefit-sharing: if it sides with the poorer nations, the convention may move towards a new treaty.



Snow business? Researchers hope to find valuable organisms in the icy Antarctic environment.

National park plan delayed

The US government started thinking about sharing the benefits of biodiversity after a famous, serendipitous discovery in the geysers of Yellowstone National Park.

In 1997, Yellowstone officials signed an agreement with the Diversa Corp., a company based in San Diego, that they hoped would lead to discoveries that returned revenue to the park. They were hoping for another *Taq* polymerase, the enzyme from Yellowstone's geysers (right) that is used in the polymerase chain reaction.

But the Edmonds Institute, a tiny environmental group based near Seattle, sued the National Park Service in 1998 over the deal, saying that it feared the commercial exploitation of the park. In March 1999, a judge refused to end the Diversa agreement, but ordered the

park service to develop a plan — called an environmental impact statement — that would serve as a framework for such benefit-sharing at all national parks. Diversa's deal was put on hold.

Five years later, the park service has yet to publish a benefit-sharing plan. Park officials say they hope to have a draft published this spring, and a final version complete by the end of the year. Meanwhile, the parks can't share in any benefits from bioprospecting on their land.

Entomologist John Varley, director of the Yellowstone Center for Resources and one of the park-service officials responsible for producing the plan, blamed the delay on the fact that no similar effort has previously been undertaken.

Beth Burrows, director of the Edmonds Institute, says she is "appalled" that the benefit-sharing plan isn't ready. "This really makes me



wonder about the quality of the oversight of the research agreements in the park," she says.

Eric Mathur, Diversa's vice-president of science, says it was a frustrating experience. "I see it as a lost opportunity," he says. Diversa has now shifted its search from the United States to Costa Rica and Indonesia.

Rex Dalton



Aquatic peril: common carp, one of the top fishing stocks in the Far East, face a threat from herpes.

Carp virus crisis prompts moves to avert global spread

Helen Pearson

Fish disease experts are meeting in London this week to figure out how to contain a viral disease that is razing populations of carp.

Fish populations are periodically struck with viral diseases, but koi herpes virus, or KHV, is killing four out of every five fish that it infects, and has spread rapidly around the world. "In the past 30 years it's the worst and most rapidly spreading virus I've dealt with," says Ron Hedrick, who studies infectious diseases in fish at the University of California, Davis.

The disease threatens two important fish populations: the ornamental koi carp industry, which is worth tens of millions of dollars in Japan, and the common carp, the world's fourth most-farmed fish.

The herpes virus was first isolated in Israel in 1998, and has since been detected in ornamental koi carp in Europe, Asia and the United States. Last year it started killing large numbers of common carp in Japan. Authorities there were alerted to the problem in October, when fish began dying in Ibaraki prefecture's Lake Kasumigaura, where more than half of Japan's farmed carp are produced.

By the end of 2003, the disease had been reported in common or koi carp in 23 of Japan's 47 prefectures, according to Motohiko Sano of the National Research Institute of Aquaculture in Tamaki.

Experts fear that the virus could cause further economic damage if it spreads to farmed carp stocks in other countries — particularly in China, which produces three-quarters of the world's farmed carp. "We could have a situation where it reaches the wild population and potentially decimates it," says Hedrick.

The London meeting, to be held on 12 and

13 February, will bring together researchers, animal-health experts and representatives from the ornamental-fish industry.

One point for discussion will be the lack of an effective test for the virus, particularly one that can reliably detect it during its sometimes lengthy asymptomatic phase. "The industry is almost screaming for a detection method that will tell you where this virus is," says Keith Davenport, chief executive of Britain's Ornamental Aquatic Trade Association, based in Trowbridge.

Researchers attending the meeting say they also want to know where the new virus first emerged — whether it jumped into carp from another species, for example. This might help them to anticipate and prevent other new fish infections. "Finding out where these things are originating might change the way they are handled in the future," says Hedrick.

Another issue is whether international trade in ornamental fish, which is thought to have fuelled the spread of the virus, should be more tightly controlled. Although fish imported for aquaculture are subject to rigorous health inspections, regulations on the transport of ornamental fish are more hit and miss.

Some researchers say that the virus should be included in a list of certifiable diseases compiled by the World Organisation for Animal Health. Nations would then be obliged to notify the agency within 24 hours of the disease being detected.

But Davenport argues that good practice within the industry can limit the disease without the need for further formal regulations. "Industry has no wish to trade in fish with this virus present," he says. ■

Kofi Annan backs call for science push in developing countries

Declan Butler

Scientific academies worldwide are calling for the establishment of two funds to boost research efforts in poor countries.

The InterAcademy Council — a coalition of 90 academies, including the US National Academy of Sciences — delivered the idea last week in a report to Kofi Annan, the secretary-general of the United Nations (UN).

The report, *Inventing a Better Future: A Strategy for Building Worldwide Capacities in Science and Technology*, slams governments for failing to get behind laboratories, institutions and policies in developing countries that could serve as powerful engines for their economic development.

It recommends that a Global Institutional Fund be set up to support 20 centres of national or regional excellence over periods of between five and ten years, selected on the basis of the quality of their science, independence, management calibre and relevance to the needs of their regions. A second Global Program Fund would operate a competitive grant system, with international referees assessing proposed joint projects between laboratories in rich and poor countries.

The funds would be supported by contributions from governments, foundations and existing international organizations, although the scale of their support is not specified.

Annan, who chaired a debate to discuss the recommendations at the UN headquarters in New York, welcomed the report and said he would work with its authors to implement their proposals.

He said his top priority was the application of science and technology in agriculture: "I challenge the scientific community to ask why Africa is the only continent not to have had a green revolution."

The 161-page report features an exhaustive list of initiatives that it says would boost science and innovation in poor countries. "A lot of what it says ought to have already happened, but it's got to be restated," says Peter Bruns, vice-president for grants and special programmes at the Maryland-based Howard Hughes Medical Institute. But he cautions that initiatives must be driven by existing research institutions, not by creating additional bureaucracies. ■

Agreement close on strategy to flush out alien stowaways

London The International Maritime Organization is hosting a meeting in London this week in an attempt to stop alien species from invading seas where they're not usually found.

Ships unwittingly transport thousands of plant and animal species in their ballast tanks, which are filled with sea water at one port to provide stability and emptied at another. Species that hitchhike in the tanks can badly damage ecosystems. Ships brought the American comb jellyfish to the Black Sea, for example, causing fish populations to crash.

The organization, which is part of the United Nations, has been considering for more than 10 years a convention to ensure that ballast waters are filtered or treated. It anticipates that an agreement will be drawn up by the end of this week's conference.

Space agency grounds fractious astronauts

Moscow Fearing a bout of space rage, officials at the Russian space agency have told two astronauts that they cannot go to the International Space Station.

Russian Valery Tokarev and American Leroy Chiao had been due to travel to the space station on 19 April, but it has been hinted that the astronauts do not see eye to eye. "It's not that the crew was unprofessional or ill," a spokesperson for the agency told Reuters. "But on certain conditions it was not ready. They should get on like good friends." Squabbling astronauts would be difficult to manage onboard the space station, which hosts crews in cramped conditions for six months at a time.

With NASA's shuttle fleet grounded, Russia's Soyuz spacecraft is the only means of ferrying people to and from the space station. The pair's stint will now be carried out by NASA astronaut Michael Fincke and Russian Gennady Padalka.

Royal Society split over merits of 'popularizer'

London Fellows of the Royal Society, Britain's most prestigious scientific body, have threatened to resign following a leak that a popular science figure had been nominated for a fellowship.

Neuroscientist Susan Greenfield, who was nominated in a secret ballot in December, is professor of pharmacology and senior research fellow at the University of Oxford, but is perhaps best known for her prolific newspaper articles and television appearances. She has even posed for UK celebrity gossip magazine *Hello!*.

Scientists win right to pick over old bones

San Diego The bones of Kennewick Man should be given to scientists for study, an appeals court ruled last week. The fate of the 9,200-year-old remains has been debated for seven years, with four Native American tribes claiming the skeleton for burial (see *Nature* 407, 548; 2000).

The San Francisco court ruled on 4 February to uphold the rights of a group of archaeologists and anthropologists to study the Kennewick Man skeleton (plastic cast of skull, right), which was found in 1996 in the Columbia River at Kennewick, Washington. It is one of the oldest and most complete skeletons found in North America.

The tribes tried to block the examination on the grounds that the remains are of an ancestor. The court ruled that the evidence so far doesn't



prove that the skeleton is of a Native American. An attorney for the tribes said they are considering their legal options.

Anonymous criticisms of the nomination, which was meant to be confidential, came from Royal Society fellows who view Greenfield's research achievements as unworthy of election to Britain's oldest scientific academy. "Her science is not of the stature that would get her into the Royal Society," sniffs one fellow, who wouldn't be identified. "But she is one of the best public faces in science and one has to admire that."

In a statement, Greenfield said "it is a great pity that those who do not have the courage to identify themselves can make unsubstantiated criticisms both of my science and of my activities in public communication."



Susan Greenfield: criticisms are 'unsubstantiated'.

Poll says young scientists lack the talent for teamwork

Washington Poor interpersonal skills are hampering the careers of young researchers, according to a poll by an online panel of scientists.

The Science Advisory Board, whose goal is to improve communication between the scientific community and the outside world, surveyed 1,400 scientists and asked them whether they thought that young trainees

are well equipped to handle their new careers. Some 44% of the scientists surveyed said that researchers who are just starting out lack essential skills for resolving conflicts they might encounter at work. One-third of those surveyed said that new researchers aren't good at working in teams. And 22% said that young scientists don't know enough about negotiation techniques.

The survey did not compare young scientists with trainees in other occupations, which could indicate whether these low rates of interpersonal skill are limited to the sciences. But the board's director, Tamara Zemlo, nonetheless says that young scientists are putting a crimp in their careers early on if they lack good negotiation and communication skills. She suggests that these skills should be part of graduate and medical-school curricula.

South Africa's science minister quits his post

Cape Town The South African politician who oversaw the country's readmission into the international science arena after years of isolation during the apartheid era has resigned.

Ben Ngubane has held the cabinet position of minister of arts, culture, science and technology since the advent of democratic government in 1994, except for two-and-a-half years between 1996 and 1999, when he was premier of the province of KwaZulu-Natal.

Ngubane's resignation has caused concern among the scientific community in South Africa, as he has been a strong advocate of increased government spending on research. He will not be replaced until after the forthcoming general election on 14 April. Phumzile Mlambo-Ngcuka, minister for mineral and energy affairs, will act as a caretaker minister. Ngubane is set to take up a diplomatic post as ambassador to Japan.



Eat lead: toxins from explosives can reach dangerously high levels at the firing ranges where soldiers and police are trained.

Collateral damage

Even munitions that are never used in anger can have a long-term impact on the environment, and the military is anxious to minimize the risks. Jim Giles talks to the chemists who are developing 'green' explosives.

“I know ... some people think it is an oxymoron.” Ron Jones sounds weary. He has had to deal with my incredulous line of questioning before. “But we really do need green explosives.”

Bombing and shelling can't be good for any ecosystem. But away from the heat of battle, military officials are ploughing precious research funds into developing less-toxic explosives that should be better both for the personnel who have to handle them, and for the environment.

Jones and his colleagues at the US Naval Air Warfare Center in China Lake, California, for instance, are trying to replace the lead-based compounds used to fire guns. Other groups want to phase out explosives whose residues not only pollute the environment when detonated, but also cause problems when unused ordnance is disposed of. “The environmental issues with weapons go beyond what you see in combat,” says Jones.

Pollution is a particularly pressing issue at the firing ranges where soldiers and law-enforcement officers are trained.

When you pull the trigger of a firearm, two small explosions follow. First a hammer or an electrical current detonates a small amount of ‘primer’ explosive. This ignites a larger amount of explosive, which forces the bullet down the barrel.

The problem is that the most widely used primers contain lead. Two of the current favourites — lead azide and lead styphnate — are responsible for the dangerously high levels of lead found at some firing ranges. A 1991 survey, for instance, found that employees who had just cleaned a range run by the FBI in Quantico, Virginia, had levels of lead in their blood almost ten times as high as US government health limits¹.

Target practice

Jones and his colleagues hope to replace lead compounds with nanoparticles of aluminium, which react violently with oxygen. They mix particles averaging 50 nanometres across with acetylene black, a form of carbon, and molybdenum trioxide. The former makes the mixture electrically conductive, allowing it to be ignited by passing a current through it. The latter supplies oxygen.

The researchers are now running tests to assess ‘action times’ — the gap between the ignition current switching on and the bullet leaving the muzzle of the gun, which for military purposes needs to be less than 4 milliseconds. If the aluminium particles pass this and other assessments, they could be in use sometime in the next decade.

At Ludwig-Maximilians University in Munich, Germany, chemists led by Thomas

Klapötke have investigated about 50 substances that could make safer primers, funded largely by the German military. A primer must detonate relatively easily to ensure that guns can fire quickly, so the researchers first measure how much energy it takes to detonate them. They do this by rubbing samples between sandpaper, dropping weights on them and spraying them with sparks.

Next up are tests that measure destructive power. About 10 grams of the test substance is placed in a standard steel container, then ignited. Each container has a hole of a particular size in the lid. The smaller the hole, the more likely it is that gases produced by the explosion will be unable to escape quickly enough, blowing the container apart. The ideal primer — which only needs to be powerful enough to ignite the secondary explosive — will destroy a 35-millilitre container with a hole in the lid 4 millimetres across.

In five years of testing only five compounds, all of them rich in nitrogen, have made it through this process. One candidate, known as TNTA, consists of a series of N_3 and NO_2 groups attached to a ring of carbon atoms². Like other nitrogen-based explosives, part of its energy release comes from single or double bonds between nitrogen atoms that break apart and re-form as more stable triple bonds. When mixed with an



Fire power: the destructive force of US weaponry is plain for all to see, but the environmental effects take a little longer to kick in.

oxygen-rich material such as ammonium nitrate, TNTA explodes to produce harmless molecules of nitrogen and carbon dioxide. The compound is undergoing commercial tests, and Klapötke hopes that it will be on the market in two to five years.

Bomb disposal squad

In addition to certain types of primer, chemists would like to see other explosives phased out. Soil and vegetation at disused military testing ranges across the United States are contaminated with unreacted TNT, which the US Environmental Protection Agency classifies as a possible human carcinogen. TNT is now only rarely used by the US military, but it is still widespread in mining, raising fears that it could pollute underground water sources.

Replacements for TNT have been in use since the Second World War, but these were chosen mainly because they are more powerful. James Short, a staff specialist

for defence laboratories in the US Department of Defense, says that the US military has millions of kilograms of explosives that have reached the end of their shelf-life of about 20–50 years, but which are difficult to dispose of because of environmental concerns.

The explosives involved — nitrogen and oxygen-containing substances called HMX and RDX — are mixed with polymers so that they can easily be moulded into various shapes for different applications. The polymer–explosive mix could be disposed of by burning, which does not necessarily raise it to a temperature at which it would explode.

Short argues that the environmental risks are not large. But communities living close to military facilities have blocked such moves, fearing that burning the polymers could release pollutants such as carbon monoxide. Military officials have an added incentive to replace HMX and RDX: they want explosives that are cheap and powerful, yet detonate less readily and so are easier to handle.

In the 1990s, Short was part of a team at the Office of Naval Research in Arlington, Virginia, that investigated alternatives to conventional explosive–polymer mixes. The group found that HMX and RDX could be mixed with polymers that melt when heated, allowing the explosives to be removed and stored for future use. Short's team also studied a new explosive, a nitrogen-rich compound called TNAZ, which can itself be melted and moulded³.

But money was an issue. It costs just a few tens of dollars to produce a kilogram of HMX or RDX, but about \$200 to create the same amount of TNAZ. The new polymers, meanwhile, are extremely viscous when melted, requiring expensive manufacturing equipment to mix in the explosives during production.

Short still believes that both options are economically viable when you consider the costs of disposing of HMX and RDX. "But it was hard to convince the officials responsible for acquiring weapons," he says. "They'd rather pay less now."

Work on green explosives continues at other US military facilities, such as the Army Research Laboratory in Adelphi, Maryland. Officials are reluctant to discuss these studies for security reasons. But hints about the type of compounds under consideration come from a closely related effort: the search for greener alternatives to the propellants used to power military rockets and space vehicles.

During the first two min-

utes of a space shuttle launch, for instance, when the need for thrust is greatest, power is generated by burning aluminium in two solid-fuel boosters. The aluminium is supplied with oxygen using a compound called ammonium perchlorate, but environmentalists note that this means the exhaust plume contains chlorine ions — which can trigger ozone-destroying reactions and acid rain.

Space agencies are continually searching for more efficient propellants and, although they believe that the environmental impact of the chlorine ions is minimal, they take such considerations into account. Three main candidates — known as ADN, HNF and CL-20 — are in the frame to replace ammonium perchlorate as an oxidizer. All are chlorine-free, consisting mainly of nitrogen and oxygen.

Blast off!

Results from tests in military and civilian labs in Europe and the United States suggest that they could improve the performance of solid-fuel boosters. But introducing a new propellant is a complex business, says Octavia Frota, an expert in energetic materials at ESTEC, the European Space Agency's research and technology-development centre in Noordwijk, the Netherlands. Tests to find the right combination of fuel and oxidizer, and a compound to bind them together, are subject to stringent safety regulations. Boosters may have to be redesigned once the right combination is found, she adds.

It could take another decade for any of the three candidate oxidizers to find regular space applications. But thanks to the presence of nitrogen in these compounds, they can act as explosives as well as oxidizers. And at least one — CL-20 — is being assessed for use as an explosive by military researchers in France, Germany and Britain, says Klapötke.

Working to develop new explosives can be dangerous. "If you make one mistake it can be your last," says Klapötke. Three years ago, one researcher in his lab lost the top halves of two fingers in an explosion. "But he is still working in the field," says Klapötke.

Labs such as his can ill afford to lose staff, due to problems with recruitment. Klapötke admits that military links are off-putting for many students. And the pool of recruits is narrowed still further by his decision only to accept students from NATO countries for security reasons. Klapötke understands why some young chemists shy away from military research, but argues that his work is necessary. "We need defence, so we have to train people," he says. "But we don't want to kill them and pollute the environment."

Jim Giles is a reporter for Nature, based in London.

1. Barsan, M. E. & Miller, A. *Lead Health Hazard Evaluation* HETA Report No. 91-0346-2572 (National Institute for Occupational Safety and Health, Cincinnati, Ohio, 1996).
2. Adam, D., Karaghiosoff, K. & Klapötke, T. M. *Propell. Explos. Pyrot.* **27**, 7–11 (2001).
3. Archibald, T. G., Gilardi, R., Baum, K. & George, C. *J. Org. Chem.* **55**, 2920–2924 (1990).

"Developing new explosives can be dangerous. Three years ago, a researcher lost the top halves of two fingers in an explosion — but he's still working in the field."

The hot hand of history

We may not have known we were doing it, but humans have been changing the climate for thousands of years, a new theory suggests. Could our ancestors have saved us from an ice age? Betsy Mason investigates.

S. FERRY/GETTY



Clear cut: did deforestation 8,000 years ago set in motion a pattern of global warming that has since staved off a mini ice age?

Less than four centuries ago, life was colder in the northern latitudes than it is today. Alpine glaciers advanced, frost killed ancient orange groves in China, and Londoners held winter festivals on the frozen Thames. The Little Ice Age began around 1300 and lasted some 600 years. But according to a new theory, similar or even worse conditions might prevail today had it not been for the activities of prehistoric people.

Most scientists agree that humans have had a hand in warming Earth's climate since the industrial revolution — some even argue that we are living in a new geological epoch, dubbed the Anthropocene. But a respected climate scientist is now claiming that our involvement in climate change began thousands of years earlier and has led to more than twice as much warming as previously estimated — enough to stave off a mini ice age.

According to Bill Ruddiman, an emeritus professor at the University of Virginia, Charlottesville, humans first began to affect the climate when clear-cutting of forests for agriculture spread from the eastern Mediterranean to Eurasia 8,000 years ago. The process boosted atmospheric levels of carbon dioxide — the main greenhouse gas — by releasing carbon stored in the trees. Then, about 5,000 years ago, southeast Asians inadvertently began

adding the greenhouse gas methane to the atmosphere when they flooded fields to grow rice, effectively creating artificial wetlands.

As a result, Ruddiman estimates that by 1800, pre-industrial deforestation and farming had increased Earth's surface temperature by an average of 0.8 °C — nearly twice the rise attributed to human activity since 1850. The difference was even greater at higher latitudes — about 2 °C — enough to stop glaciers from forming in northern Canada and to render London and New York more temperate.

Warm reception

Since Ruddiman described his theory in December at the American Geophysical Union's annual meeting in San Francisco and in the journal *Climatic Change*¹, some climate scientists have been warming to the idea. "I have been thinking more and more that he may be on to something," says geologist Thomas Crowley of Duke University in Durham, North Carolina.

Many of those who spoke to *Nature* said that they might have dismissed the idea out of hand had it come from another source. But Ruddiman has a reputation for careful work. "He doesn't lay an idea out before he's really, really ready," says former graduate student Maureen Raymo, now a palaeoclimatologist

at Boston University. And he has bucked conventional wisdom before. In 1989, he was the first to propose that the rise of the Himalayas 40 million years ago altered the climate of Asia², an idea considered outlandish at the time but now generally regarded as correct.

But for other climate scientists, sizeable doubts remain. They stress that, so far, the idea is based only on correlations, and that Ruddiman has yet to present hard data. The prehistoric rise in greenhouse gases could easily be the result of other factors, they say, such as subtle changes in Earth's orbit or the release of carbon by the oceans.

Ruddiman first began to suspect an early human influence on the atmosphere several years ago, when studying sample cores extracted from ice in a region of the Antarctic interior called Vostok³. The 3,500-metre-thick sheet formed over 400,000 years as snow accumulated and compacted into ice, preserving a continuous record of Earth's atmosphere in pockets of air.

In his analysis of air bubbles in the Vostok ice, Ruddiman found that atmospheric levels of methane waxed and waned every 23,000 years for the length of the ice-core record. The methane cycle seems to be linked to the 'wobble' of Earth's axis, which periodically causes warmer summers. This in turn



Ice cores from Vostok in Antarctica reveal anomalies in Earth's atmosphere 5,000–8,000 years ago.

strengthens the monsoons that drench the tropics, expanding the surface area of wetlands — which disgorge methane from decaying plant matter.

But Ruddiman was intrigued by an anomaly in the record. “At 11,000 years ago there’s a methane peak that’s exactly where it should be,” he says. “Then the methane starts going down just the way it’s supposed to. But all of a sudden, 5,000 years ago, it just turns around and heads right back up.” At the same time, the main known source of methane, the tropical wetlands, continued to dry out. “There’s no natural explanation that’s going to work if it’s breaking the natural rules, so the only thing I could conceive of is humans,” Ruddiman says.

Grains of truth?

Ruddiman found hints that human activity might have caused the reversal in the archaeology literature. As it turns out, 5,000 years ago is when the people of southeast Asia first began transplanting rice seedlings to flooded fields to protect them from weeds and dry spells. Ruddiman proposed that the rice paddies took over producing methane where the tropical bogs had left off³.

But methane is only part of the story. The CO₂ record in the Vostok ice cores, although more complex, also seems to break a trend in recent millennia. In addition to the 23,000-year cycle, levels of CO₂ in the atmosphere fluctuate on 100,000-year and 41,000-year cycles. Although the causes of this are not yet fully under-

stood, all three cycles put atmospheric CO₂ on the downswing for the past 11,000 years. Yet 8,000 years ago, CO₂ levels began to rise.

Here again, Ruddiman cites a corresponding trend in the archaeological record. Around that time, certain domesticated grains first appear in soil sediments from southern Europe, marking the spread of farming from the Near East. Lower pollen counts from sediments in China suggest widespread deforestation as farmers cleared more land. The advent of more efficient ploughs 6,000 years ago further accelerated the trend. “Most of Eurasia was deforested by the time of Christ,” says Ruddiman.

Even in historic times, there are intriguing correlations between unexplained drops in atmospheric carbon and human activity, Ruddiman points out. For instance, drops in CO₂ levels in the third to sixth centuries and again in the fourteenth century correspond to

episodes of bubonic plague that killed 25–45% of the population of Europe and western Asia. Many acres of abandoned farmland would have reverted to forest, pulling CO₂ back out of the atmosphere, according to Ruddiman. And a third CO₂ drop between 1500 and 1750 coincides with the death of about 90% of indigenous Americans from newly imported diseases.

Ruddiman is not the first to tackle the ice-core anomalies. In 1999, one of the world’s leading climate scientists, Wallace Broecker of Columbia University in New York, proposed that much of the CO₂ came from the oceans⁴. The deep sea stores a

great deal of carbon as solid calcium carbonate. Looking at deep-sea sediment cores, Broecker’s team detected a drop in carbonate concentrations that corresponded to the rise in atmospheric CO₂. That suggested the carbon came from the ocean, not burned forests.

Ruddiman argues that Broecker’s theory and other hypotheses⁵ fail to explain why the current interglacial period — the time between ice ages — is so different from the previous three, each of which saw a drop in both methane and carbon dioxide as the next ice age approached. The sudden upswing in greenhouse gases is completely unprecedented, Ruddiman argues.

But it may not be unique, says Michel Crucifix of the Hadley Centre for Climate Prediction and Research in Exeter, UK. He has studied the partial record in the Vostok ice core from the next oldest interglacial period and says that it also seems to be different. So it’s too early to rule out natural explanations such as changes in Earth’s orbit, he says.

Root cause

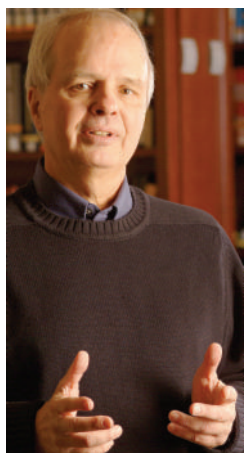
Fortunat Joos of the University of Bern in Switzerland is also unconvinced. He says that Ruddiman has underestimated the role of the oceans in absorbing carbon released through deforestation. Joos calculates that humans would have had to clear more than twice as much land as Ruddiman estimates to cause a 0.8 °C rise in temperature. He also points out that loading the atmosphere with carbon from vegetation would cause a drop in the ratio of two common carbon isotopes, carbon-13 and carbon-12. The drop seen in the ice-core data is less than half what it should be if Ruddiman is right, Joos says.

Ruddiman says that an important test of his theory will be to work out more quantitatively whether deforestation was extensive enough to cause the rise in CO₂ levels. Pollen counts from ancient lake sediments, which provide an estimate of local tree density, may offer clues. And a new Antarctic ice core that extends back about 900,000 years, recently obtained by a Swiss team, may reveal whether the current warm period is truly unique.

As grateful as we might be to our ancestors for saving us from a mini ice age — if, in fact, they did — climate scientists warn that we shouldn’t conclude that modern global warming is a good thing. We are now boosting the amount of CO₂ in the atmosphere ten times faster our ancestors did, and to levels outside the range of natural fluctuation. We may not miss winter fairs on the Thames, but no one really wants flooded coastal cities. ■

Betsy Mason recently completed an internship in Nature’s Washington DC office.

1. Ruddiman, W. F. *Clim. Change* **61**, 261–293 (2003).
2. Ruddiman, W. F. & Kutzbach, J. E. *J. Geophys. Res.* **94**, 18409–18427 (1989).
3. Ruddiman, W. F. & Thomson, J. S. *Quat. Sci. Rev.* **20**, 1769–1777 (2001).
4. Broecker, W. S. *et al. Paleoceanography* **14**, 744–752 (1999).
5. Indermühle, A. *et al. Nature* **398**, 121–126 (1999).



William Ruddiman believes that our ancient ancestors changed Earth’s climate.

Europe needs a strategy to fight kidney disease

Funding would be well spent, both in human terms and in reducing the financial burden.

Sir — Many people are unaware that autosomal dominant polycystic kidney disease (ADPKD) is one of the most common life-threatening genetic diseases on the planet, with an incidence of about 1 in 1,000. It affects some 12.5 million people worldwide, making it twice as common as multiple sclerosis and five times as common as cystic fibrosis. On top of the human cost, it causes a financial drain on health provision, including dialysis and transplantation services. In the United States the impact of ADPKD has long been recognized, yet in Europe funding for research into this disease is sorely lacking. This needs to change.

Last year, ADPKD was the only kidney disease singled out by the US Senate Committee on Appropriations for special attention. In 2003, \$1.72 billion was appropriated for the National Institute of Diabetes and Digestive and Kidney

Diseases, which has a strategic plan to tackle ADPKD and free up more than 3,000 places on kidney transplant lists, saving \$2 billion annually in ADPKD-related costs. The research budget for the major PKD charity in the United States, the PKD Foundation (www.pkdcure.org), is currently more than \$2.5 million.

The contrast with Europe is striking. In Britain, the Medical Research Council spent nothing on ADPKD in 2001–02 and just £213,000 (US\$395,000) in the previous year. Sadly, there are fewer than five specialist ADPKD research groups in Britain, even though global research has accelerated since the identification of the gene products responsible a decade ago. The UK charity the National Kidney Research Fund considers ADPKD research a strategic objective, yet it lacks sufficient high-quality applications to fund more than four ADPKD-focused projects from

a total of 80 awards. Britain's PKD Charity (www.pkdcharity.co.uk), of which I am a trustee, is in its infancy and currently has no opportunity to sponsor research.

Our European neighbours are faring no better. Budget cuts in France have hit medical research hard, including ADPKD studies, so laboratories are being depleted of their brightest young associates for lack of funding. How the next round of funding from the European Union, the Framework 6 programme, will respond to the US lead in this area is yet to be seen. Within the Framework's calls for research to tackle major disease, there is no particular focus on ADPKD, despite a clear emphasis on rare genetic conditions. Without such strategic thinking, European ADPKD research will continue to suffer.

Peter J. Lockyer

Signalling Programme, Babraham Institute, Babraham, Cambridge CB2 4AT, UK

Celebrating supernovae that changed the world

Sir — Further to J. L. Heilbron and W. F. Bynum's Commentary "1904 and all that" (*Nature* **426**, 761; 2003), two anniversaries connected with supernovae are worth noting. First, there was the discovery of the Crab supernova by some Chinese astrologers in 1054 (this would be 9.5¢ in the units used by Heilbron and Bynum). The second was the discovery of one of the renaissance supernovae, the SN 1604, by Johannes Kepler — who is mentioned in the Commentary, but for different reasons.

The Greeks thought of the sky as unchanging and did not expect to find new phenomena in the heavens. Even the comets were thought of as essentially terrestrial. However, the Chinese did not have these reservations, and kept a track of any changes in the night sky, which were believed to affect the fortunes of the country and the emperor. They have always been wonderful record keepers, and the Crab supernova is one of the many 'guest stars' they have discovered during the past 2000 years. The Crab is now known as the 'Rosetta Stone' of modern astrophysics, because its extraordinary features provide clues to numerous astrophysical phenomena, including supernova remnants, pulsars and high-energy emissions.

As your Commentary noted, Kepler and Galileo changed physics and astronomy in fundamental ways, but the general

populace could only be convinced of the need to revise overall Greek thinking about the heavens by more direct evidence — if at all. It is said that the appearance of the two renaissance supernovae (the first one found by Tycho Brahe in 1572) in the European skies helped to generate acceptance of these new scientific ideas about the heavens.

P. R. Vishwanath

Indian Institute of Astrophysics, Sarjapur Road, Koramangala, Bangalore 560 034, India

Was Watson and Crick's model truly self-evident?

Sir — I am compelled to reply to the comments of *Nature's* emeritus editor John Maddox (*Nature* **426**, 119; 2003) concerning *Nature's* editorial decision not to send the Watson and Crick paper (*Nature* **171**, 737–738; 1953) for peer review. Maddox's retrospective comment "the Crick and Watson paper could not have been refereed: its correctness is self-evident" represents the wisdom of hindsight. To suggest that any model is correct and self-evident on its face leads to acceptance without questioning.

In response to Maddox's comment that the paper could not have been refereed, several capable reviewers were available — Rosalind Franklin, Maurice Wilkins or Erwin Chargaff, for example.

Even Watson and Crick had serious doubts about the correctness of their DNA

structure. The 1953 *Nature* paper expresses self doubt: "previously published X-ray data on deoxyribose nucleic acid are insufficient for a rigorous test of our structure ... it must be regarded as unproved until it has been checked against more exact results".

In the F. L. Holmes book *Meselson, Stahl and the Replication of DNA* (Yale Univ. Press, New Haven, 2001), the author notes that Watson "suffered from periodic fears that the structure might be wrong and that he had made an ass of himself. ... Watson really did harbor serious doubts about the validity of their structure for DNA, before and after he and Crick published their first paper in *Nature*".

In a letter dated 12 March 1953 to Max Delbrück at Caltech, Watson states: "The X-ray pattern approximately agrees with the model, but since ... we have no photographs of our own ... this agreement in no way constitutes a proof of our model. We certainly are a long way from proving its correctness. ... In the next day or so Crick and I shall send a note to *Nature* proposing our structure as a possible model, at the same time emphasizing its provisional nature and the lack of proof in its favor."

Stanley Scher

PO Box 9456, Berkeley, California 94709, USA

correspondence

Contributions to Correspondence may be submitted to corres@nature.com. They should be no longer than 500 words, and ideally shorter. Published contributions are edited.

Soul searching

What have advances in neuroscience told us about the mind?

Soul Made Flesh: The Discovery of the Brain and How it Changed the World

by Carl Zimmer

Free Press: 2004. 384 pp. \$26

To be published in the UK by Heinemann in April, £17.99

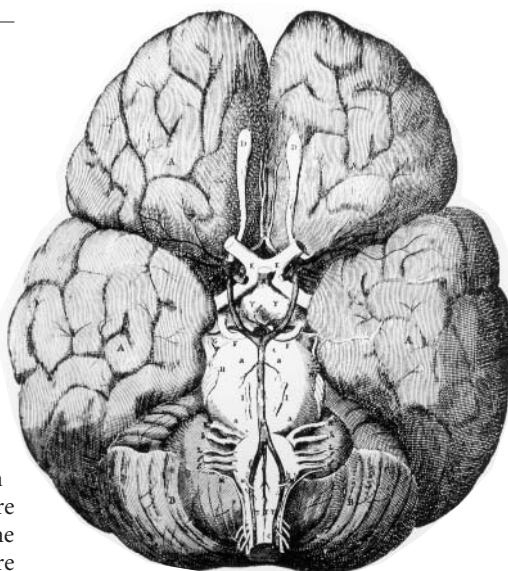
Rina Knoeff

The provocative title of this book alone, *Soul Made Flesh*, made me want to read it. After all, we live in a time where scientists are doing their best to explain all mental phenomena in terms of matter. Advanced medical technology, such as the magnetic resonance imaging (MRI) scan, makes the processes of the brain visible and gives us the feeling that we are gaining an insight into the nature of the mind and its functions. Even more, we are relieved when psychological disorders can be explained in terms of a malfunctioning brain. It enables us to ascribe our depressions and mania to something alien that is not essentially part of our true selves.

Carl Zimmer's fascinating book shows that what we think of as recent developments began in the seventeenth century, when the English anatomist Thomas Willis began dissecting brains in Beam Hall at Oxford. At the beginning of the seventeenth century, the brain was seen as a 'bowl of curds', functioning as a kind of refrigerator to cool the heat of the blood. By the end of the century, thanks to Willis, the brain was studied as the seat of emotions, perception and memory.

Surprisingly enough, Willis is largely unknown. Some might remember him as the discoverer of the so-called 'circle of Willis', a ring of blood vessels at the base of the brain, but it is not widely known that his descriptions of the brain and nerves are at the root of modern neurology. According to Zimmer, this is because John Locke's ideas eclipsed those of Willis. Locke argued that we cannot know much about the inner working of the mind, so we should restrict ourselves to the ideas themselves and how they are confirmed by everyday experience. Willis, on the other hand, showed that the anatomy and chemistry of the brain could reveal the working of the mind — an idea that was dangerous at the time because it smacked of atheism.

Yet, Willis managed to keep out of theological trouble by strictly separating the immortal, rational soul from its bodily counterpart, the sensitive soul. The latter, Willis argued, consists of tiny particles that function as messengers, conveying sense and motion from the brain via the nerves to the rest of the body, and vice versa. Illness



Christopher Wren's drawing of the brain shows blood vessels discovered by Thomas Willis.

is caused by miscommunication and damage to the brain. Most notably, Willis was able to explain psychological illnesses and disabilities resulting from brain damage.

Zimmer's book is fascinating, not least because it provides a vivid picture of the world of which Willis was a part. Not only does he describe the smells of seventeenth-century Oxford in a way that transports you right there, but he also convincingly explains Willis's work in its political and religious contexts. We meet Oliver Cromwell and his puritan followers as well as King Charles II and his curiosity for the new experimental natural philosophy. Thomas Hobbes and Lady Anne Conway also figure in the story. Last, but not least, the reader is made part of the audience watching the often outrageous and crazy experiments of the so-called Oxford experimental circle — a group of natural philosophers busy with topics as diverse as submarines, blood transfusions, spacecraft and vacuum pumps.

Historians of science have written extensively about seventeenth-century English natural philosophy. Yet making sense of the diverse experiments of the Royal Society and its forerunners remains a tricky enterprise. Historians usually focus on one or two philosophers, a series of experiments or a particular philosophical movement. Zimmer's book is remarkable in that it offers a multifaceted picture of the world and work of Willis. In so doing he has managed to make sense of Willis's ideas in turbulent times of revolution, plague and fire.

In the final chapter, Zimmer moves from the dissection room in seventeenth-century

Oxford to a present-day MRI investigation at Princeton University. He introduces us to the philosopher Joshua Green, who is interrogating someone in the machine while he looks at depictions of the brain on his computer screen. The aim of his research is to establish a better understanding of the nature of moral judgements. The soul made flesh — it is an eerie reality of modern neurological and philosophical research. To my taste, Zimmer does not sufficiently distinguish Willis's concept of soul from its twentieth-century counterpart. For Willis, the concept of soul referred to the processes of life at large, whereas nowadays we associate the soul with mental functions.

In the seventeenth century, people tried to cure diseases of the mind by manipulating the brain and nervous system. It is this continuity of concern that makes Zimmer's book such a fascinating read.

Rina Knoeff is in the Faculty of Arts and Culture, Maastricht University, 6200 MD Maastricht, the Netherlands.

Systematic survey of the spheres

Physics of the Solar System: Dynamics and Evolution, Space Physics, and Spacetime Structure

by Bruno Bertotti, Paolo Farrinella & David Vokrouhlický

Kluwer: 2003. 701 pp. \$229, £128, €209

Doug Hamilton

Progress in our understanding of the Solar System has been extremely rapid over the past decade. It has been fuelled by the discoveries of extrasolar planets, the Kuiper belt and myriad distant planetary satellites, by spacecraft fly-bys of asteroids and comets, by the *in situ* exploration of Mars, and by the advent of fast computers and efficient numerical algorithms. To the great benefit of students and practitioners alike, the appearance of advanced textbooks has been almost as rapid. The most recent offering is *Physics of the Solar System*, a well-written and comprehensive overview of the diverse bodies that surround the Sun and of the intricate interplay between them.

Physics of the Solar System is much broader in scope than two other recent advanced textbooks, *Solar System Dynamics* by Carl Murray and Stanley Dermott (Cambridge University Press, 2000) and Alessandro

Morbidelli's *Modern Celestial Mechanics* (Taylor & Francis, 2002). These titles focused primarily on orbital dynamics, a topic that accounts for only about half of *Physics of the Solar System*, which manages to cover it clearly and succinctly by outlining many derivations rather than providing full details.

Preceding the orbital dynamics is an opening section that emphasizes the physical states of planetary bodies, including their rotation, gravitational fields, tidal distortions and heat budgets, and another that focuses on magnetospheres and atmospheres. The volume concludes with several chapters devoted to issues relevant for artificial satellites. The book is more focused and deeper than *Planetary Sciences* (Cambridge University Press, 2001) by Imke de Pater and Jack Lissauer, but is less broad, containing no chemistry and only a limited discussion of geological topics.

Physics of the Solar System is tightly written, fun to read and should appeal to experts in the field and new graduate students alike. Within its covers abound a wealth of interesting and little-known nuggets of planetary lore that, although available in the scientific literature, have not appeared in an accessible text before. For instance, it is well known that planets move along elliptical paths around the Sun and that these orbits precess (or rotate) slowly in space under the influence of weak gravitational perturbations from the other planets. What is less widely known is that these elliptical orbits do not always precess independently; those of Jupiter and Uranus, for example, oscillate about one another, while precessing at a common rate.

The book has a nice introduction to planetary atmospheres that starts from the concept of hydrostatic equilibrium. The authors show that an atmosphere of constant temperature continually loses gas molecules to space, and they make an analogy to the solar wind. Magnetic fields are treated in greater depth than is usual for a planetary textbook, but without the overwhelming detail of a text on plasma physics. The authors keep their focus on the Solar System by pointing out, for instance, that the energy source that powers Earth's magnetic dynamo is thought to arise from heat released as Earth's core solidifies. The lack of ongoing core solidification in Venus, rather than the slow rotation of our sister planet, is the preferred explanation for its lack of an active dynamo.

The authors provide a clear and insightful discussion of planetary gravitational fields. They describe how the dipole terms of an expansion of a gravitational field vanish with the choice of the centre of mass as the origin, and show that two of the five quadrupole terms disappear when the z -axis is chosen to point along the spin axis. They show that although geostationary satellites drift slightly relative to a fixed point on Earth's equator,



Inspiring: Antoni Gaudí's unfinished church, the Sagrada Família, dominates the Barcelona skyline.

there are two stable longitudes where this drift vanishes, thanks to a 1:1 resonance with the geopotential. These longitudes are defined by the minimum equatorial diameter of Earth, which pierces our planet near the Maldives and the Galapagos — true islands of stability in Earth's chaotic seas.

There is also a beautiful presentation of orbital-perturbation theory. It begins with a derivation of the perturbation equations of celestial mechanics based on changes to energy and angular momentum. Subsequent discussion applies the equations to the examples of atmospheric drag and Earth's quadrupole field. I follow the same scheme when teaching my undergraduate orbital-dynamics class.

Each chapter comes with a generous set of useful and challenging problems, making the text appropriate for a graduate-level course on the Solar System. One minor quibble is that the chapter on the space-time structure of the Solar System, which is a brief introduction to the general theory of relativity, is less well linked to the book's other topics than one expects from the book's subtitle. Even so, I strongly recommend making space on your shelf and time in your schedule for this lively, interesting and authoritative volume. ■

Doug Hamilton is in the Department of Astronomy, University of Maryland, College Park, Maryland 20742-2421, USA.

Science and the city

Walks around the Scientific World of Barcelona

by Xavier Duran & Merce Piqueras
Ajuntament de Barcelona: 2003. 362 pp.
€19. www.bcn.es/publicacions

Jacqueline Reynolds
and Charles Tanford

The Spanish city of Barcelona dates back to the third century BC. It expanded greatly under Roman occupation from 27 BC to AD 14 but remained largely unaffected by the Moorish invasion that left its mark on so much of the southern Iberian peninsula. Barcelona's strategic location on the Mediterranean trade routes helped it to develop into an important industrial and commercial centre. Traditional guide books emphasize the city's beautiful architecture, much of it designed by Antoni Gaudí, along with the wealth of museums, the artistic tradition, the famous football team and the hosting of the 1992 Olympic Games. Knowledgeable travellers will probably associate the painters Pablo Picasso and Salvador Dalí with the city, but few would make any connection with science,

despite the prominence given to this field by today's media.

Barcelona's city council commissioned this beautifully produced guidebook to fill the gap between its artistic heritage and the less well known scientific and technological tradition. The citizens have a significant appreciation of scientists, and not just those associated with their city. Street names remind us of Pasteur, Archimedes, Descartes and Fleming, for example, and the hall in the Industrial School has walls covered with engravings commemorating scientists from all over the world and from many different disciplines. There is even a monument to the city's direct role in determining the standard metre. The French definition of this distance was one ten-millionth of a quadrant of the Earth on the longitudinal line passing through Paris; as part of this project, Jean Delambre and Pierre Méchain were assigned the task of measuring precisely the distance between Dunkirk and Barcelona.

This guidebook is divided into two major parts: the first describes the scientific and technological history of the city, and the second comprises suggested itineraries through the various geographical areas. There are also four excellent maps, as well as a multitude of colour illustrations. The book is so wide-ranging that it provides something for everyone. Outdoor types can use it as a guide when they view remnants of the old Roman aqueduct and a short stretch of canal that carried water to the city for several hundred years. A stroll along Notariat Street leads to the house occupied by the Nobel-prizewinning neurologist Santiago Ramón y Cajal during his tenure at the University of Barcelona. Botanical gardens of all sizes abound.

For the more traditional traveller there are museums dedicated to geology, ethnography, medical instrumentation, pharmacology and many varied types of technological applications. And when all this science begins to be a bit much, there is the Picasso Museum and the Museum of Catalan Art, both of which are highly praised by the ubiquitous Michelin Green Guide. Here one can enjoy not only the visual pleasures on display but also an entertaining discussion in this book about the relationship between the arts and science.

We have never been to Barcelona, but reviewing *Walks Around the Scientific World of Barcelona* has led us to consider making the trip from North Yorkshire to this marvellous city. The one caveat is that the book's excellent production quality has led to it being a somewhat weighty tome, difficult for those of us no longer in the bloom of youth to carry on our sightseeing tours. ■

Jacqueline Reynolds and Charles Tanford are authors of *The Scientific Traveller: A Guide to the People, Places and Institutions of Europe* (Wiley, 1992).



Up close and personal: images on giant screens oversee a virtual anatomy lesson in *Tulp, the Body Public*.

Performance

A visceral experience

Tulp, the Body Public

www.elision.org.au/projects/tulp

Carina Dennis

An anatomy lesson delivered over the flayed arm of an executed thief is the inspiration behind Justine Cooper's *Tulp, the Body Public*, an intimate and confrontational performance event that explores the theme of medical surveillance of the human body.

An Australian currently living in New York, Cooper weaves together a video collage of vox-pop narratives with fragments of baroque opera and contemporary music composed by John Rodgers and performed by ELISION, an Australian new-music ensemble. *Tulp* premiered last month at the Sydney Festival and will be performed at the Brisbane Festival on 23–25 September 2004.

Cooper collected anecdotes from members of the general public from a booth at the Art Gallery of New South Wales during the month before the performance. As a vehicle to intimacy, the volunteers had vascular ultrasound Doppler recordings taken, before being interviewed about their

attitudes to medicine, pain, life and death.

Close-up camera angles of volunteers' faces as they tell their stories capture some visceral experiences: one man's attempt at home surgery using insufficient anaesthetic, a woman's recovery from breast cancer, the torture of treatment for scoliosis, and a grandmother's experiences as a surrogate mother.

The 60-minute multimedia performance, — inspired by Rembrandt's painting *The Anatomy Lesson of Dr Nicolaes Tulp* — draws on baroque imagery, which is projected onto two giant latex screens on either side of the centre stage. The soft gurgling of the vascular ultrasound recordings is an audio backdrop to the contemporary compositions, periodically interrupted by improvisations in which the musicians 'dissect' their instruments — by analogy with medical interventions. The audio and visual stimulation is overwhelming at times, even to the point of distraction.

The audience is simultaneously watching and being watched. The soprano, who provides a link between the narrative and the music, carries a wireless camera strapped to her wrist. She sometimes directs it at the audience, projecting their images onto the central screen. ■

Carina Dennis is Nature's Australasian correspondent.

Extinction: past and present

The fossil record, together with modern data, can provide a deeper understanding of biological extinction and its consequences.

David Jablonski

Extinguishment is a fundamental part of nature — more than 99% of all species that ever lived are now extinct. Whereas the loss of ‘redundant’ species may be barely perceptible, more extensive losses of whole populations, groups of related species (clades) or those that share particular morphologies (for example, large body sizes) or functional attributes such as feeding mechanisms, can have profound effects, leading to the collapse of entire ecosystems and the extermination of great evolutionary dynasties. The challenge is to understand both the causes — particularly the biological attributes that govern species’ vulnerability — and the consequences of extinction.

These challenges are of more than academic interest. Today’s biota is beset by many stresses, such as habitat destruction and fragmentation, over-exploitation and invasive species. In addition, species are susceptible to chain reactions that can destabilize them from the top down (by removing predators and other consumers) or from the bottom up (by removing or replacing primary producers). The most daunting obstacle to assessing and responding to these problems is the lack of anything close to a full accounting of present-day biodiversity: the 1.75 million known species probably represent less than 10% of the true inventory, and the figure is surely less than 1% for genetically distinct populations. Attempts to estimate global extinction from rates of habitat loss may eventually be verified, but a more effective strategy has been to analyse extinction in groups where the (approximate) size of the species pool is known, as in North American birds, tropical palms or Australian mammals. Such analyses have generally found, first, that the extinction or endangerment of species and populations is proceeding at an alarming pace, and second, that selectivity of extinction or decline tends to match theoretical expectations. For example, species with slow population growth rates, low population densities, or narrow geographic ranges tend to be more extinction-prone.

However, empirical data on extinction risk do not always follow neat theoretical lines. For example, large body size is associated with vulnerability in primates and birds, but is unimportant in carnivores, reptiles and marine mollusks. What emerges as an important issue is the covariation among the many traits that affect extinction risk — species with high population densities tend



Loss and gain: the extinction of the dinosaurs allowed the diversification of the mammals.

to have short generation times, small body sizes and so on — so that indirect effects may underlie apparent patterns. Further, this covariation is complex, with relationships among traits often defining polygonal fields rather than linear trends (for example, small-bodied forms may be widespread or spatially restricted). And the impact of different factors may depend on the extinction mechanism, such as habitat loss versus introduced predators. We are only beginning to get to grips with such problems.

The fossil record is a spectacular archive of extinction, and provides a vital deep-time perspective on factors governing extinction patterns. Analyses of present-day biodiversity often can only measure the net outcome of past extinction and origination, whereas the fossil record is a direct window onto raw rates of taxonomic survival or extinction. Despite differences in scale and taxonomic coverage, palaeontological analyses of extinction tend to corroborate theory and present-day data, and extend them to evolutionary timescales in important ways.

The major mass extinctions are a different story. These rare events (five over the past half-billion years, each estimated to have removed more than 60% of marine species) bring not only surges in extinction intensity, but often shifts in selectivity as well. Factors that are apparently significant for ‘normal’ extinctions (such as local abundance, reproductive mode, body size, feeding strategy, geographic range at the species level and species richness) had little effect on the survival of clades during the mass extinction at the end of the Cretaceous period of 65 million years ago, and were unimportant in one or more of the other mass extinctions as well. The ‘big five’ mass extinctions appear to fit a model of ‘non-constructive selectivity’ — not strictly random, but often selecting on

features such as clade-level geographic range that are not tightly linked to traits favoured during ‘normal’ times, and thus are unlikely to reinforce or promote long-term adaptation. Perhaps intrinsic factors diminish in importance with increasing perturbation, as also seen in today’s freshwater fishes and Australian marsupials. Those shifts in selectivity, along with the sheer magnitude of the ‘big five’ events, may explain why post-extinction periods are famously important in opening opportunities for once-marginal groups, for example the expansion of mammals after the dinosaurs’ demise.

In assessing extinction and the diversity of the remaining biota it is important to look at more than just the number of species lost. This is because the random loss of taxa generally has a weak impact on morphological diversity, whereas extinctions of clades can have far greater impact. Thus, the same taxonomic extinction intensity can have different effects on the morphological variety of the survivors, and significant adaptations can be lost simply because they belonged to a taxon that lacked other features, such as broad geographic range, that promoted extinction-resistance.

We still do not know how directly relevant the ancient mass extinctions will be to the present-day situation. But even excluding the ‘big five’ events, the fossil record contains a multitude of natural experiments on, for example, global warming or cooling episodes that exceed the ability of species to cope with local conditions, the disassembly of ecological communities, the extinction of formerly robust taxa and the colonization of previously sparsely populated habitats. All of these experiments are relevant to present-day systems. The comparative calibration of extinction magnitudes and effects, and how they relate to the initial state of the system, to the nature of the driving mechanisms, and to post-extinction physical and biotic conditions, will all be important components of a fuller theory of biodiversity dynamics. ■

David Jablonski is in the Department of Geophysical Sciences, University of Chicago, 5734 South Ellis Avenue, Chicago, Illinois 60637, USA. He thanks S.M. Kidwell and M. Foote for valuable comments.

FURTHER READING

Erwin, D. H. *Proc. Natl Acad. Sci. USA* **98**, 5399–5403 (2001).

Jablonski, D. *Proc. Natl Acad. Sci. USA* **98**, 5393–5398 (2001).

Purvis, A., Jones, K. E. & Mace, G. M. *BioEssays* **22**, 1123–1133 (2000).

Lattice window on strong force

Ian Shipsey

A long-awaited breakthrough has been made in lattice quantum chromodynamics — a means of calculating the effect of the strong force between sub-atomic particles that could, ultimately, unveil new physics.

The fundamental particles called quarks exist in atom-like bound states, such as protons and neutrons, that are held together by the strong force. The heavier varieties of quark, such as the bottom quark, can disintegrate to produce other, lighter particles, and the pattern of the decay rates is constrained, but not determined, in the theory of fundamental particles, the standard model. That pattern, especially the part involving the bottom quark, is sensitive to new physical phenomena. But although accurate measurements of the rates have been made, the window on new physics has been obscured. This is because the binding effect of the strong force between quarks modifies the decay rates: unless correction factors can be accurately worked out, the data cannot be fully interpreted for signs of any physics that is as yet unknown. This has been the case for almost 40 years. But at last, Davies *et al.* report an advance in lattice quantum chromodynamics, a method of calculating the effect of the strong force, that promises the calculational precision required (C. T. H. Davies *et al.* *Phys. Rev. Lett.* **92**, 022001; 2004).

The standard model describes all observed particles and their interactions. Particles interact by exchanging other particles that convey force. For example, in an atom, electrons bind to protons by swapping photons with one another. This is the electromagnetic force, described by the theory of quantum electrodynamics (QED). In a proton, two types of quark, called 'up' and 'down', are bound together so tightly, by exchanging particles called gluons, that this is known as the strong force. Its associated theory is quantum chromodynamics, or QCD. In the standard model there is a third force, the weak force, which is the mediator of radioactive β -decay. Another example of the weak force in action is the decay of a heavy bottom quark into an up quark, through the emission of a W particle (which then itself decays to an electron and an anti-neutrino; Fig. 1a).

Despite its success, the standard model leaves many questions unanswered. For example, although the observable Universe is made of matter and there is no evidence for significant quantities of antimatter, equal amounts of both should have been created in the Big Bang. When matter and antimatter meet, they annihilate each other: if a small asymmetry between matter and antimatter did not exist at the time of the Big Bang, there

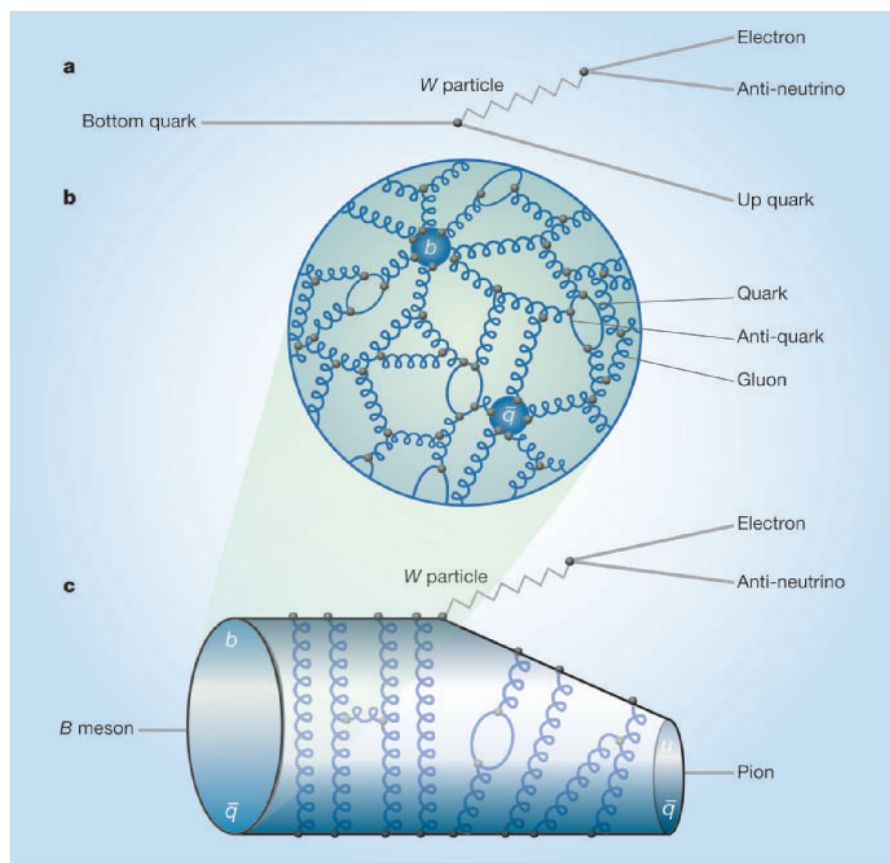


Figure 1 **Bottom's up.** a, An idealized representation of the decay of a free bottom quark into an up quark. In the standard model of particle physics, the process occurs through the weak force, mediated by a W particle, and also produces an electron and an anti-neutrino. b, In the real world, however, there is no such thing as a free quark. Instead, a bottom quark exists in a bound state with other quarks — such as in a B meson, bound by the exchange of gluons to an anti-quark. Gluons and quark pairs are constantly emitted then reabsorbed; only a fraction of this 'sea' of particles is shown here. c, So the realistic picture of the decay of a bottom quark is complex. The B meson — a bottom quark and anti-quark pair — becomes a pion (an up quark and an anti-quark), but the route is obscured by the mass of gluons and quarks (of which, again, only a fraction are shown). Calculating the details of the process is fiendishly complicated. But new advances in lattice quantum chromodynamics mean that precise theoretical correction factors can be worked out, and the problem effectively reduced to the simple process shown in a.

would be no matter in the Universe today. So how did that asymmetry arise?

If heavy particles that existed in the early Universe decayed preferentially into matter over antimatter, that could have created the matter excess. In the standard model, two types of quark, bottom and strange, do decay asymmetrically. But this effect alone is far too small to account for the asymmetry. However, there are many theories that predict the existence of other, massive particles that could readily produce the asymmetry. And

because of the connection between asymmetry and mass, these theories also address other puzzles, such as why electrons are almost 10,000 times lighter than bottom quarks.

Searching for evidence of these particles can be done directly or indirectly: powerful accelerators, reaching ever higher energies, could create these mysterious particles; or there is the precision approach of looking for subtle deviations in the properties of known particles, influenced by the unknown. If

deviations are found, their pattern and magnitude, much like a fingerprint, would identify the new physical phenomena responsible. To see such deviations in the decay rates of quarks, a measurement precision at the level of 1–2% is required, and this is becoming achievable experimentally. But the necessary correction factors, allowing for the binding effect of the strong force between quarks, have been riddled with uncertainties at the 20% level.

Consider the rate at which a bottom quark transforms into an up quark (Fig. 1a). In principle, it is an easy measurement to make: produce a known number of bottom quarks and count how many times one of them disintegrates and an up quark is produced. However, bottom quarks always come paired with lighter anti-quarks, in particles called *B* mesons (Fig. 1b). The bottom quark and the anti-quark constantly exchange gluons, within the bounds of Heisenberg's uncertainty principle, binding them strongly together. So experimenters must study the decays of *B* mesons, not free quarks (Fig. 1c). In measuring the decay rate, correction factors must be used to compensate.

The uncertainties in these correction factors are due to the great strength of the strong force. Similar corrections arise in QED, through the constant exchange of photons. But because the electromagnetic force is so much weaker than the strong force, the quantum corrections are tiny. A detailed calculation of any electromagnetic process can be performed by adding up a sufficient number of quantum corrections, and consequently the theory of QED has been verified down to the tenth decimal place.

Not so QCD, where there is an additional complication: unlike photons, the gluons have a property called colour, which means that not only do the quarks in a meson constantly exchange gluons, but those gluons can constantly exchange other gluons as well. Pairs of quarks and anti-quarks also make fleeting appearances. The quantum corrections that allow for this swarm of gluons and quarks are very large, so adding them all up is unfeasible. As a result, QCD calculations involving the strong force cannot be made as precisely as those for QED processes.

The way round this is to use powerful computers to simulate the most probable arrangements of quarks and gluons inside a particle, and from there to estimate the particle's properties. But no computer in existence could keep track of all the quarks and gluons in a meson. The problem can be simplified by imagining space and time not as a continuum, but as a lattice — a four-dimensional grid of discrete points. Quarks and gluons reside at these points. With this restriction, an infinite number of variables is reduced to a finite (although very large) number of variables. This approach is called lattice QCD.

Originally developed in the 1970s, lattice

QCD initially enjoyed great success; in the 1980s, it explained why quarks are bound inside protons. But it took almost 20 years to go from predictions of qualitative features to realistic calculations of particle properties. Now, thanks to the increasing speed of computers and, more importantly, to a succession of improvements in the technique, Davies *et al.* have managed to calculate nine different quantities, each a particle property that is determined by the strong force. They have included in their calculation all of the pairs of light quarks and anti-quarks that fluctuate into brief existence inside a quark-based particle (Fig. 1b) — previously these pairs had usually been left out, because they are much more difficult to deal with than gluons, and it has been prohibitively costly in computer time to simulate them. Each property calculated by Davies *et al.* has already been measured in experiment, and — the acid test — all of the calculations are a good match to the data, to within a few per cent.

The authors are now racing to calculate the properties of other particles called *D* mesons, which contain a heavy 'charm' quark. Experiments are already under way with the CLEO-c detector at Cornell University in New York, to measure the

properties of these particles to a precision of several per cent. If the lattice calculations are completed by the time the CLEO-c measurements are published, and they agree, it will serve as a powerful validation of lattice QCD and silence the sceptics. Such validation is crucial if similar lattice QCD calculations of correction factors for *B* mesons are to be confidently applied to data from current and future experiments around the world. The success of the lattice approach would also be relevant in other areas of particle physics, and in nuclear physics and astrophysics.

If all goes well, experimentalists and theorists together will soon pull back the strong-force curtain that has confounded them for 40 years and look through the window for the first time. If they see a pattern of quark-decay rates that does not conform to the standard model, the deviations might provide information about new physical phenomena — phenomena that make bottom quarks heavier than electrons and give rise to the asymmetry between matter and antimatter that permits us to exist. ■

Ian Shipsey is in the Department of Physics, Purdue University, West Lafayette, Indiana 47907, USA.
e-mail: shipsey@physics.purdue.edu

Medicine

Genetic spotlight on a blood defect

Diether Lambrechts and Peter Carmeliet

The causes of defects in the blood system of newborn babies can be hard to establish if the errors are not inherited. An elegant approach has identified a gene that can encourage new blood vessels to grow.

Babies who are born with defects in their vascular system face serious medical and social problems. But we know little about the cause of these blood-vessel anomalies^{1,2}. Discovering the genetic basis of such vascular birth defects remains a challenge, as most of these errors are not inherited. Instead, they occur sporadically and affect only certain areas of the body.

On page 640 of this issue, Tian *et al.*³ report the discovery of the first susceptibility gene for a disorder characterized by diverse defects in the vascular system — Klippel–Trenaunay syndrome (KTS). The authors show that when the gene, called *VG5Q*, is expressed at high levels, new blood vessels are stimulated to grow, suggesting that *VG5Q* is probably responsible for the vascular malformations seen in some patients with KTS. The unusual means by which Tian *et al.* identified and evaluated *VG5Q*, with a combination of human genetics and functional assays, underscores the importance of using similar approaches to identify other factors involved in the formation of new blood vessels (angiogenesis).

Such factors could include molecules that are clinically relevant, or potential drug targets.

Blood vessels are lined by special endothelial cells and surrounded by smooth muscle cells. They are formed when founder endothelial cells give rise to a simple network of blood vessels — a 'vascular plexus' — which is subsequently remodelled into a more mature network of large and small vessels². This remodelling allows blood to carry oxygen to growing tissues, and so is essential for fetal development. If remodelling fails to occur normally, the vessels might become deformed, resulting in vascular birth defects. Unlike blood-vessel tumours, in which endothelial cells grow in excess, vessels in vascular malformations have normal numbers of endothelial cells but are improperly formed and remodelled. A few molecules, such as vascular endothelial growth factor and the protein Tie2, have been implicated in vascular birth defects^{1,4}, but most vascular malformations remain unexplained in terms of the genes and molecules involved.

More than a century ago, the French physicians Maurice Klippel and Paul Trenaunay

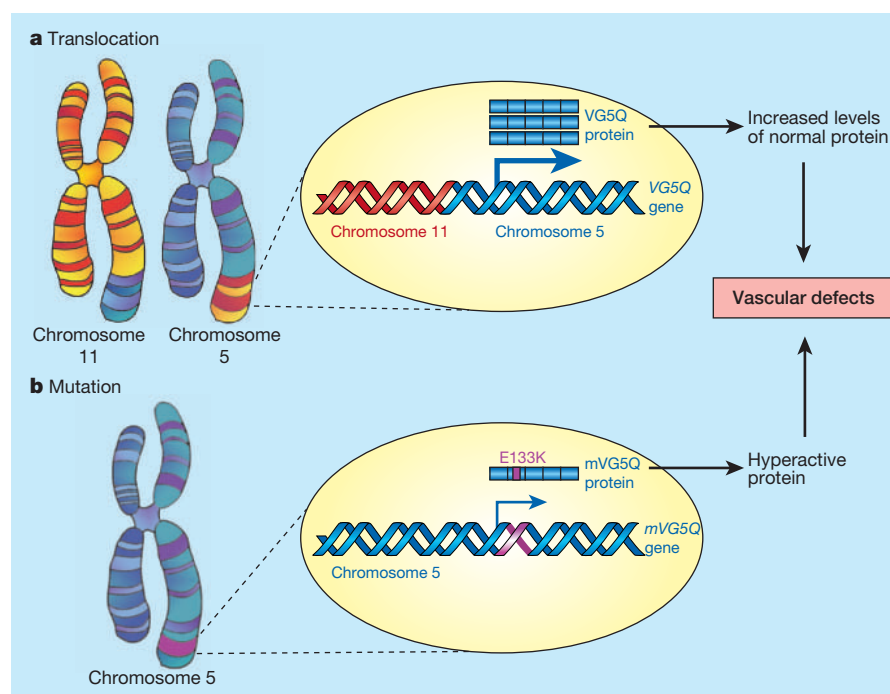


Figure 1 Alterations in the *VG5Q* gene that lead to vascular defects in Klippel–Treunaunay syndrome. **a**, As a result of a translocation between chromosomes 5 and 11, sequences of chromosome 11 become apposed to control sites in the *VG5Q* gene on chromosome 5. This increases the levels of the normal *VG5Q* protein — an angiogenic factor. **b**, An error in the coding sequence of *VG5Q* causes a mistake (E133K) in the protein. This mutant protein (m*VG5Q*) is thought to be hyperactive and therefore to increase angiogenic activity. Either mechanism increases susceptibility to the vascular defects typical of Klippel–Treunaunay syndrome.

described (and gave their names to) a rare birth-defect syndrome that is characterized by malformations in the blood vessels of the skin, often occurring only in a single, enlarged limb⁵. Patients with KTS might, in fact, have a wide range of vessel defects in various parts of the body⁶. In most cases, it is the veins, rather than the arteries, that are affected. Superficial veins in the skin are often severely twisted and swollen, probably because the deeper network of veins hasn't formed properly. In other KTS patients, however, the basic fetal network of veins persists after birth at particular locations and fails to be remodelled into a more mature network. But other vessels, including the arteries, and more frequently the lymphatic vessels (which take fluid that has leaked out of the blood vessels back into the venous system), might also be affected. Such variation in symptoms strongly suggests that KTS is caused by different genes and by many factors. But until now, genetic links to the origin of KTS were entirely lacking.

Tian *et al.*³ turned to human genetics in their hunt for the molecules that cause KTS. One child with the syndrome was reported⁷ to have a chromosomal translocation — where a chromosome breaks and part of it fuses with another chromosome — between chromosomes 5 and 11. The authors found that chromosome 5 breaks within a certain region of the *VG5Q* gene (see Fig. 1). Although this region does not contain sequences that encode the *VG5Q* protein, it

has some elements that regulate the expression of *VG5Q*. This means that when part of chromosome 11 is apposed to it, the expression of *VG5Q* increases.

But most KTS patients do not have this chromosomal translocation. In fact, Tian and colleagues found several patients who have a mutation in the coding region of the *VG5Q* gene. This results in substitution of one of the amino-acid constituents of the *VG5Q* protein by another — a glutamine (E) by a lysine (K). The mutation, known as E133K, changes the properties of the protein, and Tian *et al.* speculate that it becomes hyperactive (see Fig. 1). So increased expression of the normal version of *VG5Q*, or normal levels of the hyperactive version, could both cause blood-vessel defects consistent with KTS.

Besides the clinical link, what is the evidence that *VG5Q* actually affects angiogenesis, and how might it do so? Tian *et al.* used several methods to find more clues. First, although both endothelial and smooth muscle cells make the *VG5Q* protein, it mainly binds to and affects endothelial cells. Second, *VG5Q* promotes angiogenesis in chick embryos *in vivo*, whereas blocking its synthesis in cultured endothelial cells prevents them from forming vascular tubes. Third, the potentially hyperactive (E133K) version of *VG5Q* stimulates angiogenesis *in vitro* more strongly than does the normal protein. Finally, the angiogenic activity of *VG5Q* might, at least in part, be attributable to its interaction



100 YEARS AGO

At the close of the long frost in February, 1895, strange phenomena occurred in connection with Lough Neagh, in the north of Ireland... The lake had been frozen over for a fortnight, and thousands of people had indulged in skating on ice almost as smooth as glass. On February 22, the last day but one of the skating, though unknown to the multitudes gathered near Antrim, the ice in the central portion of the bay broke up, but left intact a sheet of about a third of a mile wide along the south-eastern shore. At a point about six miles from Antrim, this unbroken shore portion was at intervals of a few yards for a mile and upwards raised into little tunnels or bridges, from beneath which pieces of ice, large and small, along with some boulder stones of considerable size, were shot on to the land, eventually forming a ridge varying in height from two to fourteen feet, and perhaps twenty feet broad at the base. The jingling and crashing heard during the operation, which lasted for two days, were very great, and to some persons residing near most alarming.

From *Nature* 11 February 1904.

50 YEARS AGO

With the ever-increasing amount of effort being devoted to research in nuclear physics, it is perhaps no longer surprising to learn of the production of yet another artificial element — number 99 in the Periodic System (*The Times*, February 3, quoting a statement from the U.S. Atomic Energy Commission)... The nucleus 99^{247} is understood to decay by alpha-particle emission with a half-life of a few minutes. Although this achievement is not unexpected, it reflects the greatest credit on Prof. G. T. Seaborg, of Berkeley, California, and his collaborators. It will be interesting to learn what name the discoverers will propose for the element. The chemical properties may be predicted with some certainty on the basis of Seaborg's actinide hypothesis, and if such properties are in fact observed, one may refer to the element temporarily as ekaholmium. The rare earth holmium was named after Stockholm. One wonders whether the discoverers of element 99 can find a name for it, based on a place-name, which is anything like euphonious.

From *Nature* 13 February 1954.

[Editorial note: element 99 is called einsteinium, symbol Es.]

with TWEAK, a member of a family of proteins that also induce angiogenesis *in vivo*⁸. So there are data supporting a role for VG5Q in angiogenesis, although the precise molecular mechanisms involved require elucidation by more extensive *in vivo* experiments.

So is the mystery of the genetic basis for KTS solved? Certainly not. The chromosomal translocation has been found in only one affected child, and the E133K mutation is present in less than 4% of KTS patients. Many more KTS susceptibility genes are probably also involved. We do not know why the distribution of these blood-vessel defects is usually focal and varied, when all the patient's cells have the same genetic make-up. Perhaps variations in the VG5Q gene might not all be manifest, and only lead to vessel defects in particular tissues through other factors that have yet to be discovered. Alternatively, variation in the VG5Q gene might, by itself, be insufficient to produce defective vascular formation, but instead only increase the susceptibility of patients to develop KTS — that is, a second 'hit', a mutation of either the residual normal VG5Q gene or another gene, must then occur. As the latter mutation would not affect all the cells in the body, focal defects would develop only at

particular locations⁹. This is not without precedent — a second hit in the gene for glomulin probably explains the focal nature of other vascular malformations¹.

What are the overall implications of Tian and colleagues' findings²? First, they demonstrate the usefulness of genetic screening, which should prompt vascular biologists, clinicians and geneticists to consider similar strategies. Second, they show that mutations in VG5Q can disrupt angiogenesis in humans, which points to VG5Q as having the potential — hopefully, even a therapeutic one — to modulate angiogenesis. ■

Diether Lambrechts and Peter Carmeliet are at the Centre for Transgene Technology and Gene Therapy, Flanders Interuniversity Institute for Biotechnology, University of Leuven, B-3000 Leuven, Belgium.
e-mail: peter.carmeliet@med.kuleuven.ac.be

1. Brouillard, P. & Vikkula, M. *Clin. Genet.* **63**, 340–351 (2003).
2. Carmeliet, P. *Nature Med.* **9**, 653–660 (2003).
3. Tian, X.-L. et al. *Nature* **427**, 640–645 (2004).
4. Stalmans, I. et al. *Nature Med.* **9**, 173–182 (2003).
5. Klippel, M. & Treunay, P. *Arch. Gén. Méd. Paris* **3**, 611–672 (1900).
6. Berry, S. A. et al. *Am. J. Med. Genet.* **79**, 319–326 (1998).
7. Whelan, A. J. et al. *Am. J. Med. Genet.* **59**, 492–494 (1995).
8. Kolschoten, G. M. et al. *Biochem. Pharmacol.* **66**, 1427–1432 (2003).
9. Steijlen, P. M. & van Steensel, M. A. *Am. J. Med. Genet.* **85**, 359–360 (1999).

Plant ecology

Favoured aliens for the future

Peter D. Moore

Some species of plant will prefer a world with higher levels of atmospheric carbon dioxide. When those species are invasive pests, the invaders may well flourish at the expense of the native vegetation.

Concern about the rising levels of carbon dioxide in the atmosphere centres on global warming. But CO₂ is of course also the raw material of photosynthesis, and the growth of some plants is often limited by its relatively low ambient concentration (less than 0.04% by volume in the atmosphere). Such plants may be stimulated to grow by any increase in the gas, while others are not, resulting in a shift in the competitive balance in a plant community.

This is a worrisome prospect for conservationists, given the possibility of non-native, invasive plants being favoured at the expense of a native flora. An example of such a situation is described in *Functional Ecology* by Hättenschwiler and Körner¹. They have studied the response of the invasive cherry laurel (*Prunus laurocerasus*) to elevated CO₂, and compared it with the responses of the native shrubs and trees in the forests of Switzerland.

As is the case with many invasive plant pests, the cherry laurel has been transported by humans away from its native region, in this case from around the Black Sea in southwest Asia. It has been introduced into many parts

of Europe, where it is used as a decorative garden shrub or as a hedge. Cherry laurel provides dense, evergreen cover, and its scented flowers and black, bird-dispersed fruits have a certain appeal. But the cyanide-generating, toxic properties of its leaves, so familiar to entomologists, ensure that it is well protected from vertebrate and invertebrate grazers. As a result it competes vigorously against other plants that grow in its favoured habitat of shady forest floors. Like *Rhododendron ponticum*², an evergreen shrub from the same region of Asia, it has become an aggressive invader in many of the regions of Europe where it has been introduced, including the deciduous forests of Switzerland³.

Several broad-leaved evergreen shrubs from more southern climes are currently proving invasive in Switzerland, possibly assisted by climatic changes that have led to milder winters and wetter summers. But it could be that rising atmospheric concentrations of CO₂ are also having a direct influence on changing competitive balances among the plants of Swiss woodlands. Hättenschwiler and Körner examined this possibility by

constructing open-topped chambers within the forest in which germinating seeds of cherry laurel were planted along with seedlings of native plants, including holly (*Ilex aquifolium*), ivy (*Hedera helix*), hornbeam (*Carpinus betulus*) and ash (*Fraxinus excelsior*). Three treatments were provided, one in which ambient CO₂ levels of 365 μmol mol⁻¹ (p.p.m.) were maintained within the chambers, together with one treatment of 500 p.p.m. and another of 660 p.p.m.; the plots were then allowed to develop over three growing seasons. Despite the depredations of mice, 24 chambers survived the experiments and were harvested to supply data on relative growth rates.

Cherry laurel seedlings that had been exposed to increased levels of CO₂ showed a 56% increase in biomass accumulation over the three years compared with the control, although there was no significant difference between the two enhanced CO₂ treatments. Holly, on the other hand, showed no significant increase in biomass when given additional CO₂. Ivy doubled its biomass under raised CO₂, and increased its length of stem growth by 137% at 660 p.p.m., but the hornbeam seedlings were unaffected by CO₂ level. The strong differential responses between species suggest that the changing atmospheric concentration of CO₂ is likely to have a substantial impact on community composition in the Swiss deciduous forests. Not only will the invasive alien cherry laurel be favoured, but the native climbing ivy may also flourish under the new conditions, which could in turn affect tree growth — ivy weighs heavily on tree canopies and can render them susceptible to wind damage. Frost-sensitive ivy is close to its climatic limit in these woods, where daily mean temperatures in January are -2 °C, so increasingly mild winters will also favour its growth.

In his seminal work on invasive species, published in 1958, Charles Elton⁴ wrote of an approaching ecological explosion of plant and animal pests as a consequence of increasing movements of goods and people around the world. His prophecies have proved correct, but the further influence of global change could fuel the explosion to levels beyond all expectation. Field experiments exploring the direct effects of atmospheric CO₂ concentration on invasive plants within natural communities are scarce⁵: a more intensive and internationally planned programme of research is called for. ■

Peter D. Moore is in the Division of Life Sciences, King's College, Franklin-Wilkins Building, 150 Stamford Street, London SE1 9NN, UK.
e-mail: peter.moore@kcl.ac.uk

1. Hättenschwiler, S. & Körner, C. *Funct. Ecol.* **17**, 778–785 (2003).
2. Cronk, Q. C. B. & Fuller, J. L. *Plant Invaders: The Threat to Natural Ecosystems* (Earthscan, London, 2001).
3. Hegg, O., Béguin, C. & Zoller, H. *Atlas schutzwürdiger Vegetationstypen der Schweiz* (BUWAL, Bern, 1993).
4. Elton, C. S. *The Ecology of Invasions by Animals and Plants* (Methuen, London, 1958).
5. Dukes, J. S. & Mooney, H. A. *Trends Ecol. Evol.* **14**, 135–139 (1999).

Planetary science

Magnetic Mercury

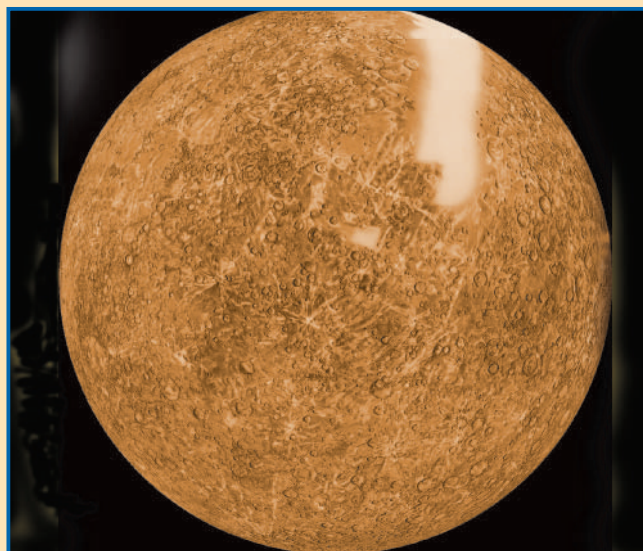
This beautiful picture of Mercury is compiled from images of the planet sent back by the Mariner 10 spacecraft in 1974 and 1975. The Mariner 10 data also presented several puzzles. Writing in *Earth and Planetary Science Letters* (218, 261–268; 2004), Oded Aharonson and colleagues revisit one of them — why, against expectations, does Mercury have a global magnetic field?

The planet's diminutive size means that it should have cooled quickly after it formed. Any molten core would have become solid, or almost completely so. A magnetic-field-generating dynamo in an outer core (as is the case for Earth) should have long since seized up. But the news from Mariner 10 meant a rethink was needed, and since then attempts to account for the magnetic field have centred on how the core

might have remained largely molten.

Aharonson *et al.* examine another possibility — that the field originates from magnetization of Mercury's crust, the remnant of a powerful core dynamo early in the planet's history. After considering some of the arguments against this possibility, the authors tackle the main objection, which comes in the form of Runcorn's theorem. This holds that a uniform shell, magnetized by an internal source, cannot have an external field if the source is removed.

Aha! thought Aharonson *et al.*, what if the shell — the crust — is not uniform, and the magnetized layer varies in thickness? They discuss why this could be so, work through the relevant mathematics, and conclude that Mercury's magnetic field could be crustal in



origin under certain conditions.

Core or crust? The question should be answered by the MESSENGER mission, due for launch in May. The spacecraft will make two passes by Mercury in 2007 and 2008, before going into orbit around

the planet in 2009. On board are a laser altimeter and a magnetometer, which can respectively provide evidence of the planet's structure and any asymmetries in the magnetic field. The resulting data should give a definitive answer. **Tim Lincoln**

MARINER 10/ASTROGEOLOGY TEAM/US GEOL. SURV.

Device physics

The optical age of silicon

Graham T. Reed

The silicon chip has been the mainstay of the electronics industry and it may similarly come to dominate photonics. A key component — a high-frequency optical modulator — has now been fabricated.

Research into optical circuits for communications began in the 1970s. Early visions of optical circuits were as 'optical superchips', containing light emitters, modulators, amplifiers, optical isolators, detectors and, latterly, electronic intelligence¹. However, there are still differing views about the optimum material for such components. This is sometimes articulated in the phrase "optical circuits have yet to find their silicon" — a reference to silicon's dominance in the microelectronics industry. Despite its success as an electronic material, silicon has received rather modest attention as an optical material. But that may be about to change. As they report on page 615 of this issue, Liu and colleagues² of the Intel Corporation have now reached a milestone that bodes well for silicon's optical future: they have fabricated the first silicon-based optical modulator with a bandwidth that exceeds 1 gigahertz (GHz).

In technology terms, research into silicon as an optical material has been under way for a long time — since the mid-1980s. However, relatively little progress has been made in comparison with more exotic materials such

as III–V compounds (indium phosphide, gallium arsenide and related compounds), the insulator lithium niobate, or even silica (which typically uses silicon as a substrate material). That is not to say that little has been achieved, but the global technical effort has largely been concentrated on materials other than silicon.

There are two main reasons for this. First, silicon does not have an inherent mechanism for the emission of light: it is an indirect bandgap material, which means that its crystal structure makes it impossible to fabricate an efficient light-emitting device, such as a laser, from this material in the conventional way. The III–V compounds, in contrast, are direct bandgap materials and for this reason are often used in semiconductor lasers.

Second, silicon does not exhibit an electro-optic effect known as the Pockels effect, the traditional characteristic for fast modulation of light (that is, encoding data onto light by selectively changing its intensity). In other materials, such as lithium niobate, modulation is typically achieved via the Pockels effect, which causes a linear change in a material's refractive index with an

applied electric field. Other optical modulators use electric-field effects known as electro-absorption and electro-refraction, but these are weak in silicon. Instead, modulation in silicon is achieved through thermal mechanisms, which are relatively slow (typically kilohertz), or through the introduction of free carriers (electrons or their positively charged counterparts, holes), which in turn results in both absorption and a change in refractive index. This latter mechanism, known as the free-carrier dispersion effect, is still also relatively slow, as it is associated with the physical movement of charge within the device. Nevertheless, it has been predicted that silicon-based modulators using this effect could achieve bandwidths exceeding 1 GHz (for example, ref. 3), although in practice working devices^{4,5} have been limited to about 20 MHz.

Recently, however, there have been dramatic changes in the fortunes of silicon-based optoelectronics. Examples of progress towards a silicon light source are: doping with erbium⁶ to overcome the indirect bandgap in silicon; building small structures that enhance useful quantum-like effects^{7,8}; and novel approaches such as dislocation engineering⁹ and optical amplification through the Raman effect¹⁰, which could result in not only an optical amplifier but also eventually an optically pumped light source. Recently, the ST Microelectronics company has announced the imminent arrival of commercial devices based on silicon light-emitting devices.

If we add to this Liu and colleagues'

announcement² of an effective silicon modulator, the difficulties surrounding the optical performance of silicon seem to be being demolished one by one. Their device is based on the free-carrier dispersion effect and has many similarities to a CMOS transistor. It comprises a slab of 'n-type' crystalline silicon (n-type meaning that it is doped to have an excess of negative charge carriers, electrons), with an upper 'rib' of p-type polysilicon (crystallized amorphous silicon with excess positive charge carriers, or holes). A thin insulating oxide layer separates the p-type and n-type regions (Fig. 1 on page 616). When a positive voltage is applied to the polysilicon, charge carriers accumulate at the oxide interface, altering the refractive-index distribution in the device — which will in turn induce a phase shift in an optical wave propagating through the device. This phase shift can be used to implement optical modulation. Furthermore, the response is fast: Liu *et al.* have demonstrated modulation frequencies in excess of 1 GHz with this device.

Of course, optical-system designers already require devices with bandwidths much greater than 1 GHz, but the demolition of this barrier by Liu *et al.* is significant: it is now easier to foresee the fabrication of even faster silicon-based modulators, although the technical challenge is still substantial. This could well signal the graduation of silicon from a subject primarily of academic research to a career as a viable photonic material with commercial applications. It is worth noting the affiliation of Liu and colleagues to the technology giant Intel — if Intel and other semiconductor technology companies can develop silicon as successfully optically as they have electronically, then silicon is certainly set to grow in stature as an optical material.

Furthermore, consider the fabrication infrastructure that exists worldwide for silicon processing: the legacy of the microelectronics industry is enormous. Those working with microelectronic devices are now seeking to move beyond the critical dimension of 90 nm that is the state of the art. However, most optical devices made from silicon are likely to have critical dimensions of hundreds of nanometres for the foreseeable future, which means that the infrastructure already exists for the next several generations of optical circuits, if they are fabricated in silicon. This implies that the economies of scale that we have seen for the electronics industry could one day apply to the photonics industry.

The removal of the 1-GHz barrier and the imminent arrival of the silicon light-emitter are not the only reasons for this change in the fortunes of silicon photonics: for some years silicon has been seen as an ideal substrate material for optical circuits. Its well-studied properties are excellent for a 'silicon bench', an accurately machined silicon layer on which components can be positioned, aligned and

attached to the outside world. Furthermore, for the networks of the future, it will be a relatively simple task to add electronic intelligence to a silicon optical device. This cannot be said for any other photonic platform.

At present the communications industry is experiencing its most serious economic downturn so far; the need for cost-effective technology solutions is underlined more strongly than ever before. Breakthrough developments such as this fast silicon modulator of Liu *et al.*² suggest that a low-cost silicon optical superchip could soon be a reality.

So let us again consider whether "optical circuits have yet to find their silicon". It is already the world's favourite electronic material, and silicon could yet come to

dominate the photonics industry as well. ■

Graham T. Reed is in the Advanced Technology Institute, University of Surrey, Guildford, Surrey GU2 7XH, UK.

e-mail: g.reed@surrey.ac.uk

1. Soref, R. A. *Proc. IEEE* **81**, 1687–1706 (1993).
2. Liu, A. *et al. Nature* **427**, 615–618 (2004).
3. Png, C. E., Reed, G. T., Atta, R. M. H., Ensell, G. J. & Evans, A. G. R. *Proc. SPIE* **4997**, 190–197 (2003).
4. Tang, C. K. & Reed, G. T. *Electron. Lett.* **31**, 451–452 (1995).
5. Dainesi, P. *et al. IEEE Photon. Technol. Lett.* **12**, 660–662 (2000).
6. Franzò, G., Coffa, S., Priolo, F. & Spinella, C. *J. Appl. Phys.* **81**, 2784–2793 (1997).
7. Canham, L. T. *Appl. Phys. Lett.* **57**, 1046–1048 (1990).
8. Pavesi, L., Dal Negro, L., Mazzoleni, G., Franzò, G. & Priolo, F. *Nature* **408**, 440–444 (2000).
9. Ng, W. L. *et al. Nature* **410**, 192–194 (2001).
10. Claps, R., Dimitropoulos, D., Raghunathan, V., Han, Y. & Jalali, B. *Optics Express* **11**, 1731–1739 (2003).

Conservation biology

Fatal medicine for vultures

Robert Risebrough

In an echo of events that unfolded earlier in the West, declines of vulture populations in the Indian subcontinent are linked to an environmental poison. Three species of these birds approach extinction.

An unusually high death rate among three species of vulture in south Asia has been perplexing scientists for several years. The consequences have been severe, not just for the vultures themselves but also for their supporting natural and human ecosystems. Vultures are natural scavengers, cleaning up carrion of wildlife and domestic livestock, and also of cattle, sacred to Hindus, which cannot be consumed by people. Moreover, to avoid contamination of the earth, water and fire, people of the Parsi faith, descendants of the Zoroastrians of the Persian empires, have traditionally left their dead to vultures for disposal.

On page 630, Oaks *et al.*¹ offer a convincing reason for the plight of the vultures. The

birds have been poisoned by diclofenac, a widely used painkiller and anti-inflammatory drug that is also an all-purpose veterinary medicine for domestic livestock in the Indian subcontinent. Diclofenac is toxic to vultures that consume the carcasses of treated livestock.

The unusual pattern of mortality was first observed in early 1997 among the Oriental white-backed vultures, *Gyps bengalensis* (Fig. 1), nesting in Keoladeo National Park in northwestern India. Birds dropped from their perches to die shortly afterwards; many remained suspended in branches until their carcasses were finally dispersed by the wind. The number of breeding pairs in the park dropped from 150 in the 1996–97 nesting

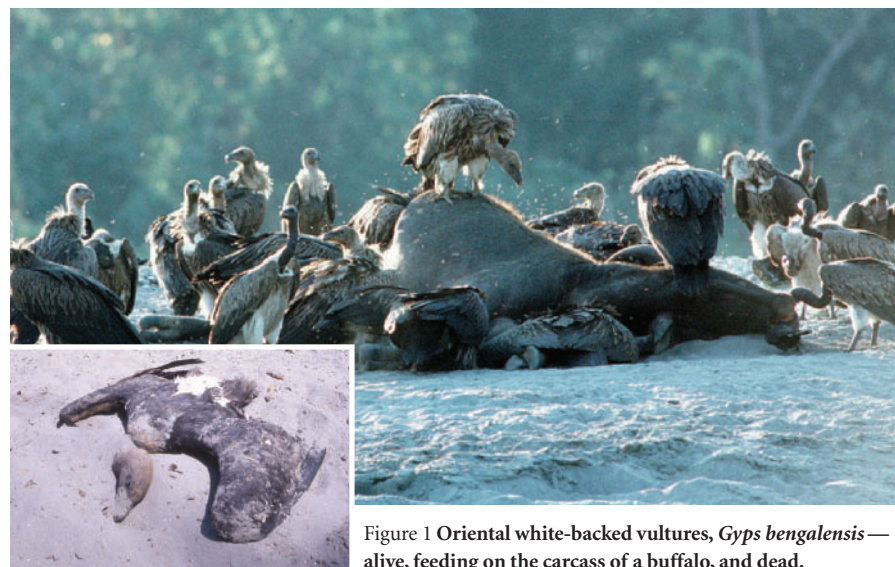


Figure 1 Oriental white-backed vultures, *Gyps bengalensis* — alive, feeding on the carcass of a buffalo, and dead.

Supramolecular chemistry

Molecular merry-go-round

The ingenuity of chemists has produced molecules that mimic a range of mechanical devices, from rotors and shuttles to switches, turnstiles and ratchets. Hiraoka, Shiro and Shionoya have now created a molecular ball-bearing, an assemblage of two disc-shaped molecules that rotate relative to each other on silver ions (Ag^+) sandwiched between them (*J. Am. Chem. Soc.* **126**, 1214–1218; 2004).

As shown in the figure, the disc-shaped molecules consist of a central benzene ring carrying either alternating thiazolyl and *p*-tolyl rings (top) or six thiazolyl rings (bottom). Mixing solutions of these molecules with silver ions results in the spontaneous formation of the ball-bearing complex. This is most stable when each silver ion is linearly coordinated to two nitrogen atoms, one each from a thiazolyl group in the

upper and lower molecules.

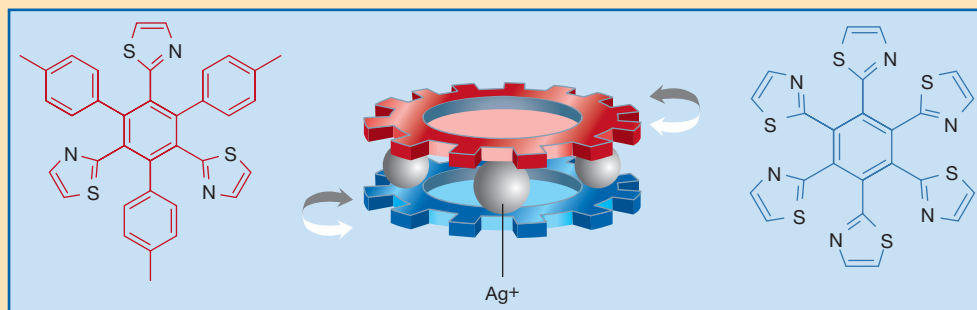
Trigonal coordination is also possible, between a silver ion and two neighbouring thiazolyl nitrogen atoms in the lower molecule and one thiazolyl nitrogen atom in the upper molecule. The complex can shift from linear coordination to trigonal to linear again, as each silver ion exchanges its original partner on the lower molecule for a nitrogen atom on a neighbouring thiazolyl group of

the same molecule. During each exchange, the two discs rotate through 60° relative to each other.

Hiraoka *et al.* use NMR spectroscopy to watch the ball-bearing in action. Close to room temperature, the complex is in full swing, like a spinning carousel, and the coordination exchanges are faster than the timescale of NMR response. But if the temperature is lowered, the coordination exchange,

and hence the rate of rotation, slows down. Then the NMR data clearly reveal that the six thiazolyl groups of the lower molecule experience two distinct chemical environments, with half of the groups coordinated to silver ions whereas the other half remain uncoordinated.

The task now is to control the speed and direction of the rotation — and put the device to practical use. **Magdalena Helmer**



season to 25, 20 and 0 over the following years^{2,3}. Anecdotal reports began to circulate of local declines of white-backed vulture populations throughout India and also of a related species, the long-billed vulture *Gyps indicus*, which occurs in peninsular India and in southeast Pakistan. In consequence, a conference was convened in Mumbai in August 1999 by Asad Rahmani, director of the Bombay Natural History Society, to consider the possible causes. In the following year, numbers of the slender-billed vulture *Gyps tenuirostris* were also found to be declining rapidly over its range in northern and north-east India. East of the subcontinent, relict populations of white-backed and slender-billed vultures survive in northern Cambodia, but very little is known of their status.

The meeting at Mumbai had a parallel with events 34 years earlier. Then, reports of the local disappearance of peregrine falcons, *Falco peregrinus*, prompted a conference that was held in Madison, Wisconsin, at the behest of J. J. Hickey⁴. Here, too, explanations for the population declines were proposed, but none could gain adequate support. At the Madison meeting, Derek Ratcliffe described his observations in Great Britain of peregrine falcons eating their own broken eggs⁴. In a now-classic paper⁵, published in 1967, he demonstrated that a physical change had begun in about 1947 in the form of reduced weights of the eggshells of peregrines and of several other species of raptorial bird. Only after an additional two years of intensive research could it be concluded that a new

environmental contaminant in the form of the DDT-derived compound DDE (dichlorodiphenyldichloroethylene) was responsible for both the shell thinning of peregrine falcon eggs in Britain and the extinction of breeding peregrines in the eastern United States.

After a comparable time from the 1999 Mumbai conference, Oaks *et al.*¹ demonstrate that the collapse of vulture populations in Pakistan has also been caused by a new environmental poison. Diclofenac is a very different kind of toxin from DDT, and travels along a very different pathway. But these findings are likely to have an equivalent impact to that produced by Ratcliffe's 1967 paper.

The Peregrine Fund, based in Boise, Idaho, began investigations of the vulture mortalities in Pakistan in 2000, in collaboration with the Ornithological Society of Pakistan. Parallel investigations were under way in India by the Bombay Natural History Society, in collaboration with the UK Royal Society for the Protection of Birds. Both groups initially focused on the search for a new disease factor, possibly one that had 'jumped' from another species; a disease hypothesis had remained the only plausible explanation for the observed mortalities^{6,7}.

In both countries, necropsies of birds that had died suddenly in good physical condition showed that they had suffered from visceral gout — the accumulation of uric acid throughout the body cavity following kidney malfunction. Kidney failure has been an

infrequent but serious side effect of the non-steroidal anti-inflammatory drugs that have come into widespread use over the past 30 years^{8–10}. That the vulture mortalities might be caused by a veterinary medicine is not a hypothesis that comes easily from a wildlife biologist. But it emerged more naturally from two of the veterinarians working with the Peregrine Fund in Pakistan, Lindsay Oaks and Martin Gilbert, who put it to the test.

An initial survey indicated that diclofenac, introduced as a human medicine in the 1970s as a painkiller and anti-inflammatory agent, had recently come into widespread use in Pakistan as a veterinary medicine. Oaks *et al.* found that kidneys of vultures dying with the visceral gout syndrome contained residues of diclofenac, whereas vultures dying of other, known, causes did not. They also found that doses prescribed for domestic animals, and meat from domestic animals treated with such doses, were lethal to captive vultures, which died with the same symptoms as observed in wild birds. Oaks *et al.* are careful to limit their conclusions to Pakistan. But diclofenac is widely used as a veterinary medicine in India¹¹, where populations of all three species continue to decline; the conclusions apply also to Nepal and Bangladesh.

Has this discovery come in time to save the three species from extinction in the subcontinent? Banning most uses of DDT permitted the recovery of peregrine falcons, but adequate, although more expensive,

substitute pesticides were available. Most, if not all, of the potential substitutes for diclofenac are related drugs¹¹ that act in a similar way^{12–14} and are also known to affect kidney function^{8–10,14}. Settling on an acceptable substitute is very unlikely to happen before the birds are extinct in the wild; moreover, ways of implementing and enforcing a ban on the use of diclofenac as a veterinary medicine must be considered in any conservation strategy. India alone has an estimated 20,000 pharmaceutical companies and 500,000 pharmacists, who take “competition to a questionable extreme”¹⁵.

The California condor, *Gymnogyps californianus*, also being poisoned by an environmental contaminant in the form of fragments of lead ammunition, was rescued from extinction by bringing all surviving birds into captivity. Such a programme, which would permit vultures to survive until they or their progeny might be reintroduced into the wild, is being discussed in India, and could yet be implemented in Pakistan. But talk has not resulted in action. In India, as occurred earlier in California, the argument is being made that ‘nature’ should be allowed to take its course. The immediate need to quarantine a sufficient number of each species to ensure their survival is ignored in discussions that focus on longer-term captive breeding and who would undertake it.

The surviving vultures are increasingly difficult to find, and because of the abundance of carcasses in the countryside, are increasingly difficult to trap. An effective conservation programme that might yet permit the survival of these birds in south Asia must emerge from the talking stage within the next several months. In demonstrating the cause of the mortalities, Oaks *et al.* also provide the impetus for the development of such a conservation programme; the timing of their paper is indeed propitious. ■

Robert Risebrough is at the Bodega Bay Institute, Berkeley, California 94705, USA.

e-mail: pelecanus@igc.org

- Oaks, J. L. *et al.* *Nature* **427**, 630–633 (2004).
- Prakash, V. J. *Bom. Nat. Hist. Soc.* **96**, 365–378 (1999).
- Prakash, V. *et al.* *Biol. Conserv.* **109**, 381–390 (2003).
- Hickey, J. J. (ed.) *Peregrine Falcon Populations: Their Biology and Decline* (Univ. Wisconsin Press, Madison, 1969).
- Ratcliffe, D. A. *Nature* **215**, 208–210 (1967).
- Cunningham, A. A. *et al.* *Anim. Conserv.* **6**, 189–197 (2003).
- Risebrough, R. W., Virani, M. Z., Katzner, T. E. & Duckworth, J. W. *Acta Zool. Sinica* (in the press).
- Schwarz, A. *et al.* *Am. J. Nephrol.* **8**, 410–416 (1988).
- Murray, M. D. & Brater, D. C. *Prog. Drug Res.* **49**, 155–171 (1997).
- Brater, D. C. *Am. J. Med.* **107**, 65S–70S; discussion 70S–71S (1999).
- Anon. *Curr. Ind. Vet. Index* **3** (Paico Press, Kochi, 2002).
- O’Brien, W. M. *Am. J. Med.* **80**, 70–80 (1986).
- Larousse, C. & Veyrac, G. *Therapie* **55**, 21–28 (2000).
- Papaioannides, D. *et al.* *Int. Urol. Nephrol.* **33**, 609–611 (2001).
- Pearl, D. & Stecklow, S. *Wall Street J.* (16 August 2001).

This article was published online on 28 January 2004 (doi:10.1038/nature02365).

came from information gleaned from preliminary structural data and the conformation of the protein⁴. Fromme *et al.* show that MutY interacts at several sites within the A·oxoG pair. But it cannot ‘flip out’ the oxoG residue from the double-stranded DNA helix. Instead, the unmodified adenine is flipped out and excised by MutY.

MutY belongs to a group of enzymes known as DNA glycosylases, which recognize altered bases in DNA and help to remove them. Like other DNA glycosylases, it generates a sharp bend in the DNA at the site of the mismatch. The new structural data provide a suitable explanation for why — as is desired — MutY doesn’t recognize and remove an adenine opposite its normal base partner, thymine (T): the extensive and precise contacts between MutY and an A·oxoG pair are entirely absent in a normal A·T pair. Similarly, the enzyme’s active site does not accommodate a cytosine opposite an oxoG; for coding reasons, it is important that the oxidized base rather than the normal base is repaired in this partnership.

In humans, different forms (polymorphisms) of the *MYH* gene have been detected that result in the production of enzymes with a reduced ability to specifically and efficiently recognize these rare A·oxoG pairs. Given that reactive oxygen species are cancer-causing, and that mutations in *MYH* are risk factors for colorectal cancer^{5,6}, the results of Fromme *et al.*³ will help in providing a detailed molecular picture of the consequences of such mutations^{6,7}. Better relative risk estimates for the development of colorectal cancer associated with a malfunctioning MYH enzyme should also gradually become available for defects that affect different sites in the protein. ■

Tomas Lindahl is at the Cancer Research UK London Research Institute, Clare Hall Laboratories, South Mimms, Hertfordshire EN6 3LD, UK.

e-mail: Tomas.Lindahl@cancer.org.uk

- Kasai, H. & Nishimura, S. *Nucleic Acids Res.* **12**, 2137–2145 (1984).
- Michaels, M. L., Cruz, C., Grollman, A. P. & Miller, J. H. *Proc. Natl. Acad. Sci. USA* **89**, 7022–7025 (1992).
- Fromme, J. C. *et al.* *Nature* **427**, 652–656 (2004).
- Guan, Y. *et al.* *Nature Struct. Biol.* **5**, 1058–1064 (1998).
- Al-Tassan, N. *et al.* *Nature Genet.* **30**, 227–232 (2002).
- Sieber, O. M. *et al.* *N. Engl. J. Med.* **348**, 791–799 (2003).
- Chmiel, N. H., Livingston, A. L. & David, S. S. *J. Mol. Biol.* **327**, 431–443 (2003).

Correction

In Karim Nader’s News and Views article “Neuroscience: Re-recording human memories” (*Nature* **425**, 571–572; 2003) the citation of reference 6 was unclear. The reference concerned — Siegel, J. M. *Science* **294**, 1058–1063 (2001) — provides a critical review of the evidence for the role of sleep in memory consolidation. Contrary to an implication of the News and Views citation, however, the author of the review concludes that there is only tenuous evidence for a connection between sleep and memory consolidation.

Molecular biology

Ensuring error-free DNA repair

Tomas Lindahl

Damaged DNA must be removed with the utmost precision, as mistakes are costly. The structure of a repair enzyme bound to its substrate provides a welcome clue to how this is achieved.

Certain forms of oxygen, known as reactive oxygen species, can be deleterious for living organisms. Although most cells can tolerate and even exploit them, these oxygen forms may contribute to cancer, tissue degeneration and ageing. In the nucleus of cells, active oxygen damages DNA, so it is important to repair an altered DNA base before it is copied during DNA replication. Failing this, the newly incorporated base opposite the original lesion may need to be removed and corrected.

One example of such repair kicks in after the conversion of guanine (G) — one of the four bases — to 8-hydroxyguanine (also known as 8-oxoguanine; oxoG) by active oxygen¹ and misincorporation of the normal base adenine (A) opposite the oxoG. A specific enzyme, called MutY in bacteria and MYH in higher organisms, breaks the link between such a misplaced adenine and the DNA sugar–phosphate backbone to initiate

repair². But the enzyme leaves well alone when oxoG is correctly paired with the base cytosine (C). On page 652, Fromme *et al.*³ describe an ingenious way by which this specificity is achieved (see Fig. 1 of their paper for a scheme of the repair cycle).

The catalytic domain of the MutY protein had already been crystallized, so the three-dimensional structure of the protein was largely known⁴. But attempts failed to co-crystallize MutY and a short piece of double-stranded DNA containing an oxoG residue and its mismatched partner, adenine, apparently because the complex was unstable.

Fromme *et al.* have overcome this technical problem by trapping the protein and the DNA together by means of a covalent bond, a crosslink. The strategy used was to generate a short stretch of DNA that had a base containing a sulphur (thiol) residue located near the oxoG. This allowed a disulphide bond to form between the oligonucleotide and MutY. Ideas for where to position the thiol residue

Human disease

Unfair sisterly exchange

Hum. Mol. Genet. **13**, 417–428 (2004)

DiGeorge syndrome, a rare condition characterized by symptoms including facial and heart abnormalities, arises through the spontaneous loss of a segment of chromosome 22 when it is passed from parent to child. Sulagna C. Saitta *et al.* show that this segment is lost in an unexpected way.

They compared copies of chromosome 22 from children with DiGeorge syndrome with those of their parents and grandparents. In 19 out of 20 families, the missing segment was lost when eggs or sperm were formed during meiotic cell division, because the resulting two 'sister' chromosomes had misaligned and incorrectly exchanged genetic information.

This is surprising — only a handful of the 20 families would be predicted to undergo such chromosomal crossovers in this particular region. It also differs from other deletion syndromes, such as Williams syndrome, in which a segment is often deleted from a single chromosome without the involvement of its homologue.

For DiGeorge syndrome, the blame may lie with large duplicated segments of DNA that are present in the deletion region, the team suspects. They hope to find out whether critical sequences might predict whether a person's chromosome 22 is susceptible to this disease-causing deletion.

Helen Pearson

Data storage

Hot bits are smaller

Appl. Phys. Lett. **84**, 810–812 (2004)

The size of individual information-storing bits in magnetic recording can now be made smaller than the accepted limit for the traditional technology. Hendrik F. Hamann *et al.* report bit sizes of less than 40 nm across in a magnetic film. This permits storage densities of 400 billion bits (400 Gbit) per square inch, and further increases of more than twofold seem possible. In contrast, the conventional technology was expected to plateau at around 150 Gbit per square inch.

The problem is that smaller bits require a finer grain size for the magnetic recording medium, but if the grains are too small they can no longer maintain their magnetic order in the face of thermal fluctuations. One answer is to use materials in which the magnetic spins are harder to flip, which are said to have higher coercivity. This makes them more resistant to thermal scrambling, but they are then also harder to write in the first place with conventional recording

heads. To record magnetic data, Hamann *et al.* use a high-coercivity material whose coercivity can be temporarily lowered by heating. By using the hot tip of an atomic force microscope, they can keep the heating highly localized.

Philip Ball

Animal behaviour

Bees clean up

Ethology **110**, 1–10 (2004)

Honeybees need to keep themselves clean — not least because their wings can become clogged by dirt, which hampers flying. But as anyone who has had a troublesome itch knows, some areas are more difficult to reach than others. Benjamin B. Land and Thomas D. Seeley have carried out a detailed investigation of the 'grooming invitation dance', by which one bee solicits a helping hand from another. This behaviour is distinct from the 'waggle dance' used to point fellow workers towards food.

Land and Seeley's first approach was to film bees in the hive. In the grooming dance, they found, a worker vibrates her body from side to side about four times a second, sometimes stopping to groom herself with her legs. In two-thirds of cases, another worker sprang to the grubby bee's aid and began grooming her. Such helpful behaviour was always preceded by the dance, and workers always touched the dancer with their antennae before beginning to groom, suggesting that the signal is transmitted by touch.

Bees can't reach the bases of their own wings. Is this the reason for the dance? To find out, Land and Seeley puffed chalk dust onto the wing bases of workers. Sure enough, dusty bees produced more dances than control workers puffed with air.

Michael Hopkin

Biochemistry

Flipping lipids

Org. Biomol. Chem. **2**, 214–219 (2004)

Cells are wrapped in a double layer of phospholipids. The inside of this membrane is rich in aminophospholipids such as phosphatidylserine (PS); the outer surface has more phosphatidylcholine (PC). This asymmetry is a key feature of cell plasma membranes, and a family of translocase enzymes, including those known as 'flippases', is dedicated to maintaining the correct distribution of lipids. Moving PS to the outside of this membrane acts as a flag to attract phagocytes, the assassin cells of the body. Consequently, synthetic flippases are being investigated as tools to elucidate membrane processes and as potential therapeutic agents.

Yoshihiro Sasaki and colleagues have previously reported synthetic flippases that move PC to the inner layer of the cell

membrane. Now they have made a flippase that can swap both PC and PS lipids in a 'flip-flop' reaction. From their studies on red blood cells, the authors believe that the flippase acts as a ferry, shuttling between the inner and outer membrane layers, carrying a lipid passenger on each journey. The authors also found that there was virtually no leakage from the cells, indicating that the flip-flop had not opened up the membrane to unregulated traffic.

Mark Peplow

Neuroscience

Astrocytes in action

Proc. Natl Acad. Sci. USA

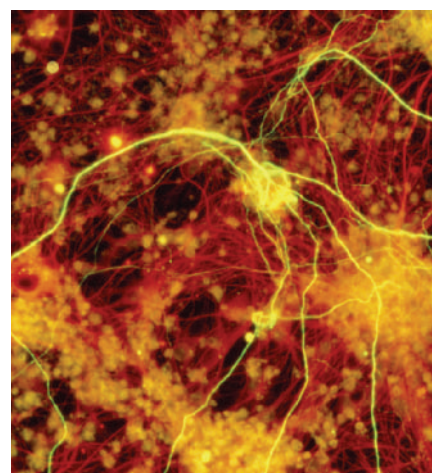
doi:10.1073/pnas.0306731101 (2004)

Glial cells were long thought of as bystanders to the main movers of brain activity, the neurons. Increasingly, however, they are taking the research spotlight. The latest developments are reported by Qing-song Liu *et al.*, who have investigated the influence of astrocytes, a type of glial cell, in the rat hippocampus.

Working with brain slices, the authors followed up evidence that, in response to intracellular rises in the levels of calcium, astrocytes release the signalling molecule glutamate. They found that increased levels of astrocyte calcium resulted in increased activity in another cell type, interneurons. Using various approaches, they came up with a chain of events in which calcium indeed prompts the production of glutamate by astrocytes, and this activates a glutamate-sensitive receptor — the kainate receptor — on interneurons.

What can be inferred about astrocyte function from these results? Interneurons have an inhibitory effect on neurons. So Liu *et al.* conclude that astrocytes may be involved in regulating the excitability level of the hippocampus, with the interneurons acting as intermediaries.

Laura Nelson



Cell sorts: a light micrograph of mammalian brain, with a neuron (yellow) set against a background of astrocytes (orange); $\times 180$.

Networking opportunity

Ian Stewart

A neglected mathematical theory is enjoying new popularity, thanks to its relevance to network dynamics in biological systems. The beating of a leech's heart is just one example that has a mathematical basis in 'groupoid theory'.

The historically uneasy relationship between biology and mathematics has improved markedly in recent years, as it has become increasingly clear that our understanding of biological systems can benefit from mathematical input and insight. Examples include the spatial and temporal behaviour of ecosystems, the geometry of protein folding, the informatics of DNA, the dynamics of the cytoskeleton and the mechanisms of visual perception. In such areas, biologists can accumulate the key data, delineate the important interactions and suggest possible mechanisms, but at some point the questions migrate to realms that are more congenial to physicists and mathematicians. Conversely, biology poses stimulating challenges for mathematics. A case in point is network dynamics. Networks arise naturally in many areas of science, and there are numerous examples in biology: gene regulation¹, protein networks², epidemiology³, ecological food webs⁴, neural networks for locomotion⁵, vision⁶, speciation^{7–9} and the synchronous flashing of fireflies¹⁰.

Many phenomena of pattern formation are common in the dynamics of networks, especially synchrony and phase-locking^{11–13}. Understanding the generalities of such patterns requires a formal theory of network dynamics to act as an intellectual framework. This stimulus from biology has encouraged a new approach¹⁴ to the problem, in terms of the mathematical theory of 'groupoids'¹⁵ — a flexible form of symmetry, which is the main subject of this article. The aim is to illustrate how biological motivation can suggest novel mathematics. Along the way, there are also glimpses of possible trade in the opposite direction.

The leech heart

Even small networks can be puzzling. A fascinating example¹⁶ — not fully explained, even by the theory to be described, but still excellent motivation — is the heart of the medicinal leech, *Hirudo medicinalis*. The heart consists of two lateral sinuses, each a tube-like series of chambers, located in the third to eighteenth segments of the leech's body (Fig. 1a). Each sinus contracts in a coordinated pattern: in one sinus, all chambers beat synchronously, in the other there is a rear-to-front travelling wave in

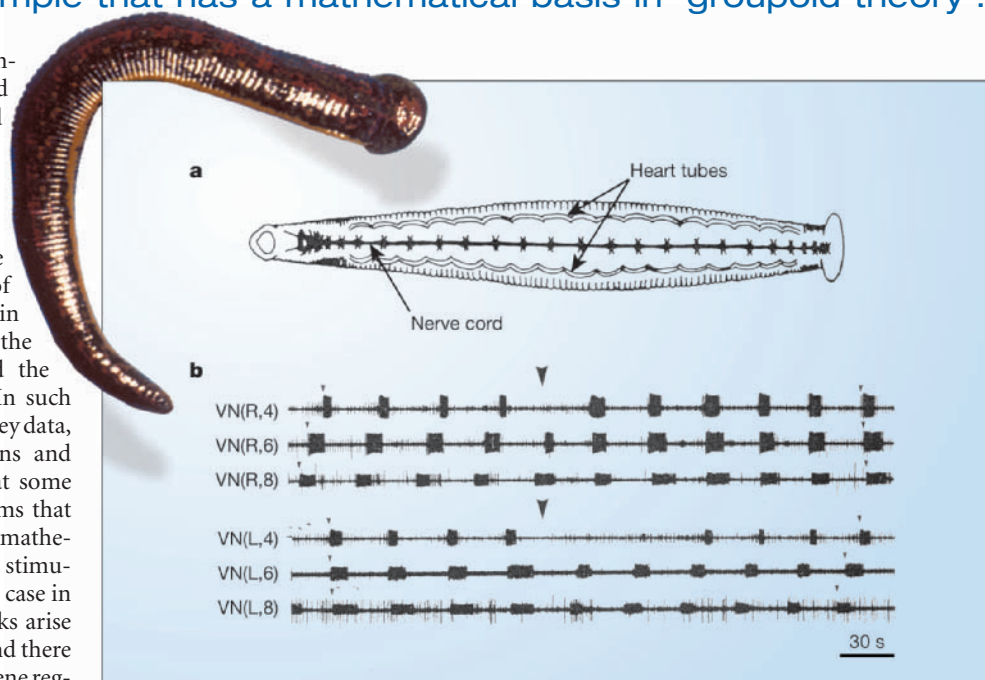


Figure 1 The heart of the leech *Hirudo medicinalis*: a, the physiology and b, the observed patterns of neural activity. Traces are shown for segments 4, 6 and 8 of the left (L) and right (R) vascular nerves (VN). The left side of the figure shows travelling waves on the right side of the heart and synchrony on the left; the right side of the figure shows the reverse, with the switch occurring at the points marked with arrowheads. The rapid 'bursting' dynamics of the neurons might be triggered by a slow dynamic with the same pattern of phase shifts¹⁸. (Graphic derived from ref. 16.)

which successive chambers contract in turn. Roughly every 50 beats, the two sides switch roles (Fig. 1b). One possible evolutionary/physiological reason for this pattern is that systolic pressure is high in the travelling wave but lower in the synchronous contractions. But what is its mathematical basis?

A mapping of the relevant neural connections has shown that this curious pattern is coordinated by a small network of neurons at the rear of the animal^{16,17}. Four of these, in segments 3 and 4, form a 'timing oscillator' that drives three pairs of interneurons (which cross-connect neural pathways) in segments 5, 6 and 7. The interneurons organize the two waveforms and control switching between them¹⁷; motor neurons transfer the timing patterns to the heart chambers. The actual signals are high-frequency bursts; these are probably driven by a slow dynamic with the same pattern of phase shifts, in which case it should be sufficient to understand the slow dynamic¹⁸.

Even though the architecture of this network of neurons has been mapped, mathematically there are unanswered questions. In

particular, it would be useful to understand how travelling waves and synchrony arise here and in networks in general, and how the network architecture produces the distinct coordinated states and switches between them. Groupoid theory goes some way towards such an understanding.

A real network has two main ingredients. First, there are 'agents', usually in profusion (for the leech heartbeat, these are various motor neurons and interneurons). Second, these agents interact, and influence one another's behaviour (for the leech, through their synaptic interconnections). The key observation that makes an abstract theory of networks possible is that the topology of those interactions — the 'architecture' of the network — can have a systematic influence on network behaviour that is more or less independent of the precise details of the agents and their interactions. An infection, for example, is transmitted more rapidly by people who have many contacts than by people who have few; with a few caveats about incubation times and the like, this is true whatever the disease might be. An associated

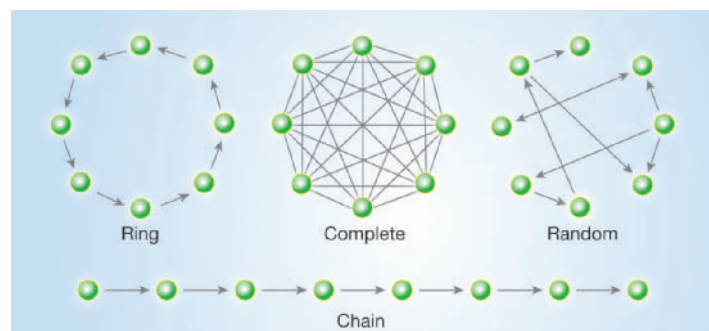


Figure 2 **Common network architectures.**

observation is the occurrence of 'motifs' — small subnetworks with specific patterns of interconnection that are unusually common^{19,20}, and may represent significant building blocks for network function.

The leech heart illustrates two important types of pattern in network behaviour: synchrony and phase-locking. The chambers along one side of the animal are synchronized — they all do the same thing at the same instant. Meanwhile, the other chambers follow the same dynamics, but shifted by some regular amount. Many notions of synchrony are found in the literature¹². Here I focus on the simplest idealization, in which two nodes are synchronous if their states are identical at all times. More realistic forms of synchrony can be viewed as perturbations of this ideal case.

Network architecture

Mathematically, a network can be idealized as a graph²¹, which consists of nodes (vertices) linked together by edges (links, connections, arrows). The nodes represent the agents, the edges represent their interactions. Two nodes are joined by an edge if, and only if, they interact. Edges may be bidirectional (interactions go both ways) or unidirectional (A influences B but not the other way round). The unidirectional case yields a so-called directed graph, whose edges are usually drawn as arrows. The edges may carry 'weights' to indicate the strength of the interaction. They may be of different kinds (fox predating on rabbit is different from rabbit predating on vegetation) or they may be nominally identical (fox A predating on rabbit X is near enough the same as fox B predating on rabbit Y). These distinctions can be taken into account by labelling the nodes and edges so that 'identical' nodes or edges carry the same label. Important architectures include chains, rings, complete graphs (all-to-all coupling) and random graphs (Fig. 2). Couplings can be to nearest neighbours, near neighbours (so many steps removed) or long-range.

We can investigate networks in the abstract (using graphs), in concrete mathematical models where the nodes and edges have additional structure, and in real networks with real agents and interactions. These three contexts are closely associated, but it is important to bear in mind that they

are different. As long as we do that, we can safely employ the same words in all contexts — for example, talking of a rabbit as a node in a food-web and drawing it as a dot.

Different people use networks for different purposes and ask different questions about them. An early pioneer was Kauffman²², who employed binary switching circuits (each node can be either on or off) to model gene interactions in a cell. He found that the dynamics of the network depended critically on the average number of edges linked to each node. The literature includes innumerable types of network: discrete (cellular automata²³), continuous (differential equations²⁴), probabilistic (Markov chains²⁵), fractal (iterated function schemes²⁶) and complex systems²⁷. A radically new network architecture that has attracted attention is the 'small world': a regular network with near-neighbour connections, in which some edges are randomly rewired to become long-range connections, or some nodes become 'hubs' that are connected to an unusually high number of other nodes²⁸.

There are several general theories of network structure and behaviour. Among them is a huge body of graph theory, invented by pure mathematicians²¹. Intensive work on the statistical properties of random graphs²⁹ has shown that, as the probability of including any given edge increases, there is a 'phase transition' at which most of the nodes suddenly link up into a single, giant component. This result shows, metaphorically at least, that if the probability of disease transmission becomes sufficiently great, almost everyone will be exposed. Less obviously, it shows that there is a sharp threshold, below which the infection remains in numerous small 'pools' isolated from one another, and above which nearly everyone is exposed.

Another approach, followed by Kuramoto³⁰, analyses network dynamics in the special case of phase oscillators with weak, linear coupling. That is, the only dynamic variable is the phase of the oscillation; the effect of any node on any other to which it is linked by an edge is small and is proportional to the state of the node or the difference between states. Think of a disease whose likelihood of being transmitted is very small, for instance. With these assumptions, useful quantitative predictions can be made about the behaviour of the network and its stability.

However, in real-world networks strong coupling and nonlinearity are rife. But these issues can be tackled. One approach, once a specific model is agreed upon, is computer simulation. An alternative is to formalize general elements of mathematical structure. In this case, the results typically apply only to special, often idealized, networks, but they can be proved with logical rigour and provide deeper understanding than just "the computer says so".

An example of general structure in a network is symmetry. Mathematically speaking, a symmetry of a network is a permutation of its nodes that preserves labels, directions and the network topology. These permutations form a group: the permutations can be inverted (still preserving the network topology); and if any two are performed in turn, the result is equivalent to performing another permutation that is a member of the group. This symmetry group of a network has a very strong influence, and the catalogue of possible behaviour is largely independent of the specific dynamics of nodes and couplings. Symmetry principles have been applied to network dynamics in locomotion⁵, vision⁶, speciation^{7–9} and many areas of physical science²⁴. However, such a concept of symmetry is too rigid for many applications. As we shall see, a more flexible view can be taken, which still preserves the structural principles of pattern formation.

Mathematical description

In the formalism¹⁴ that follows, each node determines a system of differential equations, and directed edges specify couplings. To illustrate the principles, Fig. 3 shows three simple networks: a ring, a chain and a modified chain with feedback. In these examples, all nodes are equivalent (they obey the same equations) and all couplings are equivalent (interconnected nodes affect one another in the same way). For each network there is a class of 'compatible' differential equations, describing the behaviour of each node over time. In Fig. 3a, for example, we assume that the state of any node, i , is determined by the variable x_i . (For simplicity we assume that each node has one degree of freedom, so x_i is a number.) The equations compatible with the ring network are shown in Fig. 3a, with the same function f throughout. The first variable of f represents the internal state of the node, and the second represents the node from which it receives an input.

The most striking mathematical feature of these equations is their symmetry. The graph is symmetric under cyclic permutations of the nodes, and so are the equations. The typical patterns of time-periodic states in symmetric systems are predicted by theory²⁴. For this ring network, one common pattern is a clockwise rotating wave: a periodic oscillation in which successive cells differ by a phase shift of one-third of a period. Another

possibility is an anticlockwise rotating wave (with a phase shift of two-thirds of a period).

Next, consider the chain shown, with its set of equations, in Fig. 3b. Note that node 1 has no inputs and therefore requires a distinct function g . Despite the regular appearance of the network, there is no symmetry. Each node occupies a distinct place in the chain. Indeed, with one degree of freedom for each node, the network cannot support a non-steady periodic state.

Finally, the chain in Fig. 3c is modified to include feedback. The equations contain only one function, f , because now node 1 receives an input from node 3. This network also has no symmetry; nevertheless, it supports periodic oscillations with a striking pattern. To see why, suppose that nodes 1, 4 and 7 are in synchrony, that nodes 2 and 5 are in synchrony, and that nodes 3 and 6 are in synchrony (indicated by the colours in Fig. 3c). That is, define $y_1 = x_1 = x_4 = x_7$, $y_2 = x_2 = x_5$ and $y_3 = x_3 = x_6$. The first three equations for the modified chain become:

$$\begin{aligned} dy_1/dt &= f(y_1, y_3), \\ dy_2/dt &= f(y_2, y_1), \\ \text{and } dy_3/dt &= f(y_3, y_2). \end{aligned}$$

These are the same equations as for the ring of Fig. 3a, with the x_i replaced by y_i ; the remaining four equations in Fig. 3c just repeat one or other of these equations. So, from what we know about symmetry in the simple ring network, in this network with feedback there exists a periodic state in which nodes 1, 4 and 7 are in synchrony; nodes 2 and 5 are in synchrony with each other but are one-third of a period out of phase with nodes 1, 4 and 7; and nodes 3 and 6 are in synchrony with each other but are two-thirds of a period out of phase with nodes 1, 4 and 7 (for appropriate choice of f). That is, there are travelling-wave states in which successive nodes along the chain differ only by a phase shift of one-third of a period.

This observation is striking: hitherto, travelling waves have mainly been associated with symmetric networks. The network in Fig. 3c has no symmetry, and neither do its associated equations. But when we restrict our considerations to the 'synchrony subspace' (with the assumption that $y_1 = x_1 = x_4 = x_7$, $y_2 = x_2 = x_5$ and $y_3 = x_3 = x_6$), a hidden 'rotational' symmetry becomes apparent. A travelling wave can be generated whatever the size of the feedback loop and however long the chain. Moreover, a synchronous oscillation of the entire chain arises if, instead of node 3 connecting to node 1, the first node feeds back to itself. Qualitatively, these two patterns resemble those observed in the leech heartbeat, and neurological studies^{16,17} have added important detail, as already described.

The general formalism that underlies this kind of hidden symmetry is simple but influ-

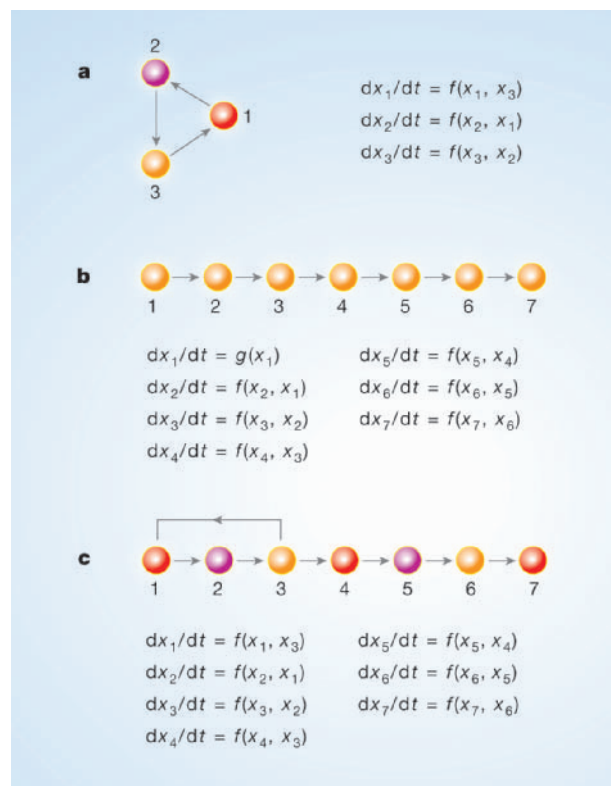


Figure 3 Describing a network. **a**, The behaviour of three identical nodes connected in this directed ring is rendered in the three differential equations shown. **b**, For a linear chain of seven identical nodes there are seven such equations. As node 1 has no inputs, its behaviour must be described with some different function, and the symmetry of the equations is broken. **c**, If feedback is introduced from node 3 to node 1, the symmetry of the equations is restored. The nodes are coloured to reflect the 'equivalence classes' that define a 'quotient network' — that is, this equivalence makes the dynamics in **c** equivalent to the dynamics in **a** (the inputs follow the pattern orange to pink, pink to yellow and yellow to orange in both).

ential. The root cause of the hidden symmetry in the modified chain network of Fig. 3c is not symmetries of the entire network, but symmetries between special subsets of the network, namely, the input sets — all nodes that emit an arrow pointing to a given node. For example, in Fig. 3a and c every node receives an input from exactly one other node. All nodes are identical and all edges are identical, so the input sets are all 'isomorphic'. In contrast, the network in Fig. 3b has two types of input structure: node 1 receives no inputs, whereas nodes 2–7 receive inputs from one other node. This difference is reflected in the appearance of a different function g for the dynamics of node 1, whereas nodes 2–7 are governed by the same function, f .

From groups to groupoids

As discussed above, the symmetries of a network — those permutations of its nodes that preserve edges, labels and directions — form a group in the usual algebraic sense³¹ if any two symmetries performed in turn produce another symmetry of the network. Input isomorphisms are similar, but they permute only a subset of nodes, namely, those whose arrows point to a given node. The corresponding algebraic structure is no longer a group but a groupoid, a concept introduced by Brandt³² in 1927. Groupoids have, on the whole, been somewhat removed from the mathematical mainstream — few undergraduate courses mention them, for example. However, they are the ideal tool for describing symmetries that apply only to parts of systems. Groupoids are more flexible and often more appropriate than the better-known

groups, and they have been unduly neglected.

To use groupoid terminology, the network in Fig. 3c has a synchrony subspace in which the dynamics is that of the network in Fig. 3a because it has a 'quotient graph': in this, the nodes are identified in three 'equivalence classes' — {1, 4, 7}, {2, 5} and {3, 6} — that respect the groupoid structure. Associated with any such quotient graph is a synchrony subspace in which states of nodes that belong to the same equivalence class are identical. The synchrony subspace is invariant under the dynamics of the network (that is, synchronized initial states remain synchronized indefinitely), and the dynamic restricted to that subspace has the form determined by the quotient graph.

Using the groupoid formalism, all robust patterns of synchrony in a network can be worked out¹⁴ — that is, patterns that occur (stably or unstably) for all differential equations compatible with the network architecture. These patterns are given by the so-called balanced equivalence relations. Informally, imagine colouring the nodes so that synchronous nodes are the same colour (as in Fig. 3c). Each colour corresponds to an equivalence class. The colouring is balanced if any two nodes with the same colour have input sets that are isomorphic via a permutation that keeps all colours fixed — for example, if any one red node has inputs from three green nodes and two yellow ones, then every red node has inputs from three green nodes and two yellow ones. Intuitively, this condition makes good sense: it means that the dynamics of any red node is given by the same function of red, green and yellow

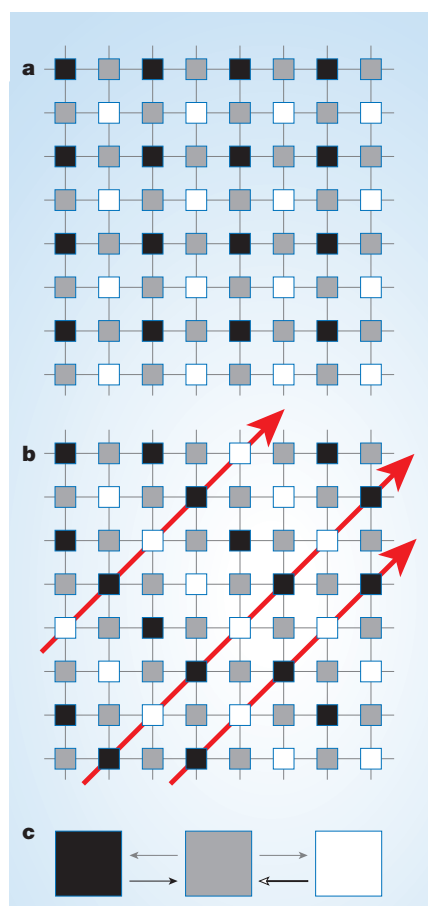


Figure 4 A square lattice, with nearest-neighbour coupling between nodes. **a**, The balanced equivalence relation is indicated by the three-colour scheme. All couplings are bidirectional, and nodes of the same colour are synchronous. **b**, Arbitrary parity changes along diagonals (red arrows, interchanging black and white squares) also produce a balanced equivalence relation. **c**, The quotient network has three nodes and is the same in all cases. Its left-right symmetry implies that there are robust, 'multirhythm', time-periodic states, in which the black and white nodes are half a period out of phase and the grey node has twice the frequency. Networks **a** and **b** therefore possess corresponding states in which identically coloured nodes are synchronous, and which can all be stable simultaneously³³.

nodes, which is a necessary and sufficient condition for synchrony to be maintained.

A more extensive example³³ is a square lattice of nodes with nearest-neighbour coupling. Here all nodes are identical, and all couplings are identical and bidirectional. The groupoid structure reveals a host of time-periodic patterns of synchrony in which three colours occur. One is highly symmetric (Fig. 4a); the rest are asymmetric (such as Fig. 4b) and can be obtained by choosing some subset of diagonal lines and interchanging the colours. However this is done, every grey node receives an input from two black and two white nodes, and every black (or white) node receives an input from

four grey ones, so the colouring is balanced. The quotient graph (Fig. 4c) is therefore the same in all cases, so all patterns can coexist in the same network. In fact, the typical periodic state here is a 'multirhythm' in which the grey nodes have twice the frequency of the black and white ones, and the latter are half a period out of phase²⁴. Thus, elements of pattern and randomness coexist in such states.

A way forward

Network architecture can impose significant constraints on the dynamics of the system, but the groupoid formalism makes it possible to approach network dynamics within a coherent theoretical framework. Just as group theory has illuminated pattern formation in symmetric systems, so groupoid theory can illuminate pattern formation in systems with repeated subunits: mathematically, the groupoid structure generalizes the group structure.

An important question is stability. Balanced equivalence relations specify robust patterns of synchrony, but 'robust' here does not equate with 'stable' in the usual dynamical sense. Further analysis, exploiting the groupoid structure, shows that the travelling waves in the chain of Fig. 3c and the patterns in the square lattice can, in fact, all occur stably for a wide range of compatible equations³³. The effect of symmetry on stability is significant and well understood²⁴; the relevant principles should generalize (though not directly) to the groupoid case. One application of such principles is food webs: although specific webs can be studied numerically, there remains a need for general ecological principles that guarantee stability⁴.

But network dynamics is far richer than merely steady and periodic states. Quasiperiodic behaviour, combining oscillations of incommensurable periods, are common. So is 'synchronized chaos', in which distinct nodes pursue identical but chaotic dynamics — a phenomenon widely studied because of its potential application in secure communications¹³ and its possible role in neural communications where certain nerve cells fire together (chaotically or regularly)³⁴. The groupoid classification of patterns of synchrony includes these more exotic types of behaviour. Subtle phenomena that occur in chaotic dynamics — such as 'bubbling', where nodes repeatedly almost synchronize only to diverge again³⁵ — probably underlie some notions of approximate synchrony.

The phenomena that are typical in networks differ in many ways from those that are typical in a general dynamical system. The network architecture not only selects particular patterns and behaviours, but permits patterns that would not normally arise at all. Philosophically, networks are a marvellous way to make exotic dynamics and patterns robust and controllable. Perhaps this is why nature makes so much use of networks.

At any rate, it places the understanding of networks high on the biomathematical agenda. The groupoid formalism is one promising route to such understanding.

Of course, specific networks governed by explicit equations can be studied through computer simulations. But in many areas, especially biology, there is no widely accepted choice of equations. In these areas — neuroscience, evolutionary biology, cell biology, ecology — general principles relating qualitative dynamics to network architecture are vital. The development of this theory will require many modelling issues to be addressed (such as approximate versus exact identity of components or states): to do so, input from both biology and mathematics will be needed. In this respect, network dynamics is a striking instance of the growing spirit of collaboration between these two major areas of science.

Ian Stewart is at the Mathematics Institute, University of Warwick, Coventry CV4 7AL, UK.
e-mail: ins@maths.warwick.ac.uk

- Guezim, N., Bottani, S., Bourguin, P. & Kepes, F. *Nature Genet.* **31**, 60–63 (2002).
- Solé, R. V. & Pastor-Satarras, R. in *Handbook of Graphs and Networks* (eds Bornholdt, S. & Schuster, H. G.) 145–167 (Wiley-VCH, Berlin, 2003).
- Grassberger, P. *Math. Biosci.* **63**, 157–172 (1983).
- Neutel, A.-M., Heesterbeek, J. A. P. & de Ruiter, P. C. *Science* **96**, 1120–1123 (2002).
- Golubitsky, M., Stewart, I., Collins, J. J. & Buono, P.-L. *Nature* **401**, 693–695 (1999).
- Bressloff, P. C., Cowan, J. D., Golubitsky, M., Thomas, P. J. & Wiener, M. C. *Phil. Trans. R. Soc. Lond. B* **356**, 299–330 (2001).
- Stewart, I. & Cohen, J. in *Nonlinear Phenomena in Biological and Physical Sciences* (eds Malik, S. K., Chandrasekharan, M. K. & Pradhan, N.) 1–63 (Indian Natl Sci. Acad., New Delhi, 2000).
- Stewart, I., Elmhirst, T. & Cohen, J. in *Bifurcations, Symmetry, and Patterns* (eds Buescu, J., Castro, S. & Dias, A. P. S.) 3–54 (Birkhäuser, Basel, 2003).
- Stewart, I. *Phil. Trans. R. Soc. Lond. A* **361**, 1101–1123 (2003).
- Mirollo, R. E. & Strogatz, S. H. *SIAM J. Appl. Math.* **50**, 1645–1662 (1990).
- Pecora, L. M. & Carroll, T. L. *Phys. Rev. Lett.* **64**, 821–824 (1990).
- Boccalletti, S., Pecora, L. M. & Pelaez, A. *Phys. Rev. E* **63**, 066219 (2001).
- Wang, X. F. *Int. J. Bifurc. Chaos* **12**, 885–916 (2002).
- Stewart, I., Golubitsky, M. & Pivato, M. *SIAM J. Appl. Dyn. Syst.* (in the press).
- Weinstein, A. *Not. Am. Math. Soc.* **43**, 744–752 (1996).
- Calabrese, R. L. & Peterson, E. *Symp. Soc. Exp. Biol.* **37**, 195–221 (1983).
- Gramoll, S., Schmidt, J. & Calabrese, R. L. *J. Exp. Biol.* **186**, 157–171 (1994).
- Golubitsky, M., Josic, K. & Kaper, T. J. in *Global Analysis of Dynamical Systems* (eds Broer, H., Krauskopf, B. & Vegter, C.) 277–308 (Inst. Phys., Bristol, 2001).
- Milo, R. *et al. Science* **298**, 824–827 (2002).
- Shen-Orr, S., Milo, R., Shmoolik, M. & Alon, U. *Nature Genet.* **31**, 64–68 (2002).
- Tutte, W. T. *Graph Theory* (Addison-Wesley, Menlo Park, 1984).
- Kauffman, S. A. *J. Theor. Biol.* **22**, 437–467 (1969).
- Wolfram, S. (ed.) *Theory and Applications of Cellular Automata* (World Scientific, Singapore, 1986).
- Golubitsky, M. & Stewart, I. *The Symmetry Perspective* (Birkhäuser, Basel, 2002).
- Grimmett, G. R. & Stirzaker, D. R. *Probability and Random Processes* (Clarendon, Oxford, 1992).
- Barnsley, M. F. *Fractals Everywhere* (Academic, Boston, 1993).
- Mainzer, K. *Thinking in Complexity* (Springer, Berlin, 1994).
- Watts, D. J. & Strogatz, S. H. *Nature* **393**, 440–442 (1998).
- Newman, M. E. J. *SIAM Rev.* **45**, 167–256 (2003).
- Kuramoto, Y. *Chemical Oscillations, Waves, and Turbulence* (Springer, Berlin, 1984).
- Burn, R. P. *Groups: A Path to Geometry* (Cambridge Univ. Press, 1985).
- Brandt, H. *Math. Ann.* **96**, 360–366 (1927).
- Golubitsky, M., Nicol, M. & Stewart, I. *Some Curious Phenomena in Coupled Cell Networks* (in the press).
- Singer, W. *Neuron* **24**, 49–65 (1999).
- Ashwin, P., Buescu, J. & Stewart, I. *Nonlinearity* **9**, 703–737 (1996).

Brain asymmetry and long-term memory

Fruitflies that have structurally similar brain hemispheres forget within a matter of hours.

The asymmetrical positioning of neural structures on the left or right side of the brain in vertebrates^{1,2} and in invertebrates^{3,4} may be correlated with brain laterality, which is associated with cognitive skills⁵. But until now this has not been illustrated experimentally. Here we describe an asymmetrically positioned brain structure in the fruitfly *Drosophila* and find that the small proportion of wild-type flies that have symmetrical brains with two such structures lack a normal long-term memory, although their short-term memory is intact. Our results indicate that brain asymmetry may be required for generating or retrieving long-term memory.

Detailed inspection of the *Drosophila melanogaster* wild-type brain revealed an unknown structure present in the right hemisphere (asymmetrical body: 'AB' in Fig. 1). This asymmetrical body is round (diameter about 10 μm), expresses the neural protein fasciclin II (FasII)⁶ and is localized near the fan-shaped body, which connects the left and right hemispheres⁷.

We searched for natural exceptions to this asymmetry in the wild-type population and identified the FasII-expressing structure in both brain hemispheres in a small proportion of flies (7.6%; $n=2,550$) (Fig. 2a, b). These wild-type flies showed no other anatomical differences compared with flies whose brains were asymmetrical in this respect. In particular, the specific brain regions involved in long-term memory⁸ appear to be normal in flies with symmetrical brains (see supplementary information). The distribution of symmetrical and asymmetrical brains is similar in the wild-type strains *Canton-Special* and *Berlin*, and in males and females.

We tested the memory of flies that had a symmetrical brain. Wild-type flies were trained to associate an odour with an electric

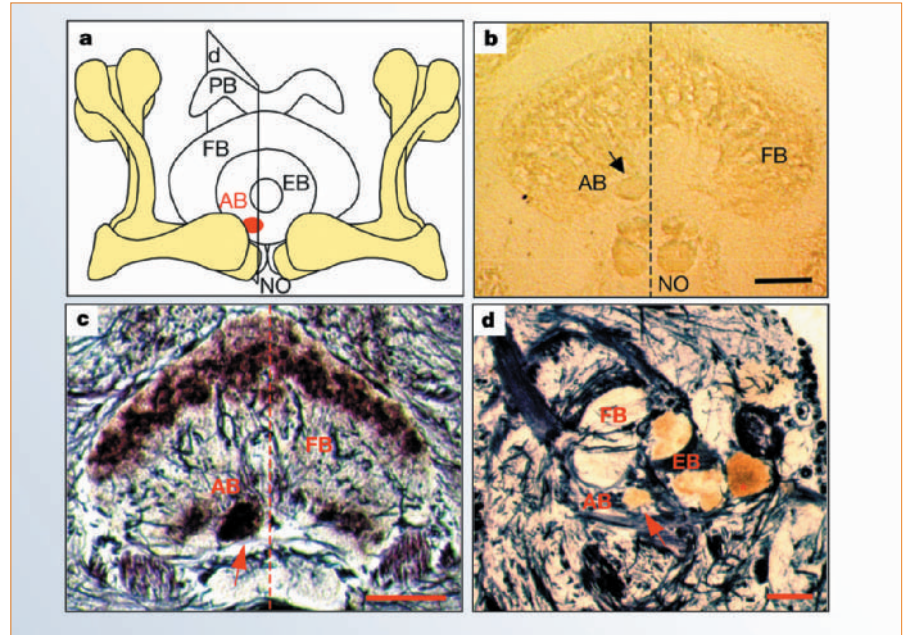


Figure 1 A structural asymmetry in the *Drosophila* brain. **a**, Central brain (dorsal, above; anterior, foreground); the sagittal-cut plane is shown ('d'). The mushroom-body lobes (yellow) and the ellipsoid body (EB), the fan-shaped body (FB), the noduli (NO) and the protocerebral bridge (PB) are shown; the asymmetrical body (AB) is in red. **b**, Frontal paraffin section, showing frontal structural brain asymmetry (arrow). **c**, Frontal section of a wild-type brain double-stained with anti-FasII antibody (brown) and Bodian staining (black and pale red). **d**, Sagittal section of wild-type brain stained as in **c**. The FasII antibody reveals the AB, the EB and the tip of the β -lobe of the mushroom body (red asterisk). Scale bar, 20 μm . Dashed lines in **b** and **c** indicate the midline.

shock in two experimental protocols: one in which a single training cycle was used to induce short-term memory⁹, and another in which intensive conditioning, consisting of five individual training sessions interspersed with 15-min rest intervals, was used to induce protein-synthesis-dependent long-term memory^{8,10}.

Flies were tested after three hours for their short-term memory, and after four days for long-term memory. Flies that made correct choices during the test were processed separately for staining with an anti-FasII antibody from those that made incorrect choices. We

then recalculated a memory-performance index post mortem for wild-type flies with asymmetrical and symmetrical brains.

Flies with symmetrical and asymmetrical brains performed comparably in the test after three hours (Fig. 2c), which indicates that brain asymmetry is not required to establish short-term memory. It also shows that flies with symmetrical brains are not defective in learning, or in odour or shock perception. However, four-day long-term memory was not evident in wild-type flies with a symmetrical brain (Fig. 2c).

Our findings indicate that structural asymmetry is important in the formation or retrieval of long-term memory in *Drosophila*. Further investigation should clarify the position of the asymmetrical body in the neural circuitry of the olfactory memory process.

Alberto Pascual*, Kai-Lian Huang, Julie Neveu, Thomas Pr  at

D  veloppement, Evolution, Plasticit   du Syst  me Nerveux, CNRS, 91190 Gif-sur-Yvette, France
e-mail: preat@iaf.cnrs-gif.fr

*Present address: Laboratorio de Investigaciones Biom  dicas, Hospitales Universitarios Virgen del Roc  , 41013 Sevilla, Spain

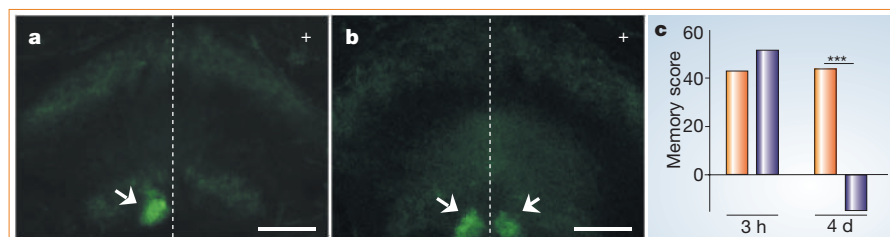


Figure 2 An asymmetrical brain is required to form or retrieve long-term memory in *Drosophila*. **a**, Single confocal section showing the asymmetrical body (arrow) in a wild-type brain, labelled by using anti-FasII antibody. **b**, In a few wild-type flies, the brain is symmetrical, presenting a double structure (arrows). Dashed lines in **a** and **b**, midline. Scale bar, 20 μm . **c**, Memory was measured three hours after a single conditioning cycle (3 hours, 1 cycle) and four days after five spaced cycles (4 days, 5 cycles). The 3-hour memory of flies with symmetrical brains (purple bars) was not significantly different from that of flies with an asymmetrical brain (orange bars; χ^2 test, $P>0.3$). However, long-term memory was significantly decreased in wild-type flies with a symmetrical brain in comparison with flies with an asymmetrical brain (χ^2 test, $P<10^{-3}$). For full methodological details, see supplementary information.

- Galaburda, A. M. *Ciba Found. Symp.* **162**, 219–233 (1991).
- Concha, M. L. & Wilson, S. W. *J. Anat.* **199**, 63–84 (2001).
- Nassel, D. R. & Strausfeld, N. J. *Cell Tissue Res.* **226**, 355–362 (1982).

4. Hobert, O., Johnston, R. J. & Chang, S. *Nature Rev. Neurosci.* **3**, 629–640 (2002).
5. Bradshaw, J. L. & Rogers, L. J. *The Evolution of Lateral Asymmetries, Language, Tool Use, and Intellect* (Academic, New York, 1993).
6. Grenningloh, G., Rehm, E. J. & Goodman, C. S. *Cell* **67**, 45–57 (1991).
7. Heisenberg, M. *Fortsch. Zool.* **39**, 30–39 (1994).
8. Pascual, A. & Pr  at, T. *Science* **294**, 1115–1117 (2001).
9. Tully, T. & Quinn, W. G. *J. Comp. Physiol. A* **157**, 263–277 (1985).
10. Tully, T. *et al. Cell* **79**, 35–47 (1994).

Competing financial interests: declared none.

Supplementary information accompanies this communication on Nature's website.

Evolutionary genetics

CCR5 mutation and plague protection

A recent and prevalent mutation in the chemokine receptor CCR5 in humans of northern European ancestry has been proposed to provide protection against bubonic plague^{1,2}. Here we infect both normal and CCR5-deficient mice with the bacterium *Yersinia pestis*, the cause of the plague epidemics that wiped out one-third of Europeans in the Middle Ages³, and find no difference in either bacterial growth or survival time between the two groups. Unless the pathogenesis of *Yersinia* infection differs markedly between mice and humans, our results indicate that CCR5 deficiency in people is unlikely to protect against plague.

A 32-base-pair deletion in the coding region of CCR5, a mutation designated as CCR5Δ32, was first identified in individuals who had been exposed to HIV but who seemed to be resistant to infection⁴. This CCR5Δ32 allele is mainly confined to caucasians^{5,6}, and its protective effect in homozygotes against transmission of the most common HIV-1 isolates arises because both CCR5 and CD4 are needed as co-receptors for entry of the virus into the cell.

Because the CCR5Δ32 allele shows evidence of very strong selection, it has been suggested that it may protect against another disease associated with high mortality^{1,2}. A candidate agent is *Yersinia pestis*, which emerged shortly after the estimated origin of the CCR5Δ32 mutation (about 800 years ago), and killed some 25 million people in the Black Death plague of 1346–52.

If plague was responsible for driving this selection, the CCR5Δ32 genotype should alter the host response to *Y. pestis* infection to improve the survival rate. But, because plague is no longer common among caucasians, the allele has not been tested for a protective effect. We therefore compared the susceptibility of two groups of mice, with and without CCR5 deficiency, to infection and death following challenge with *Yersinia*.

The homozygous CCR5Δ32 genotype is associated with intracellular retention of a truncated CCR5 protein and ablation of the chemokine response to macrophage inflammatory protein-1β (ref. 6), whereas hetero-

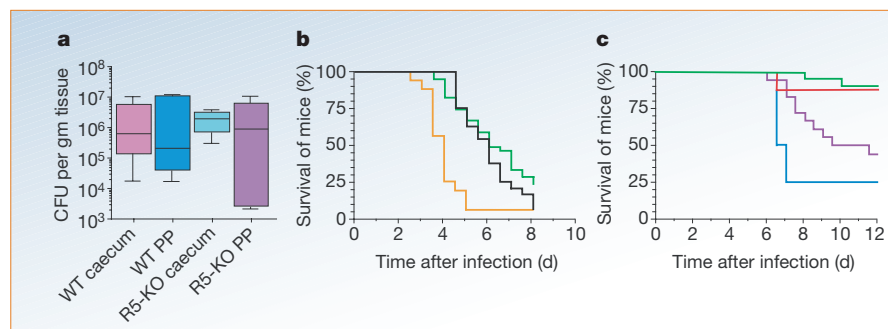


Figure 1 Impact of CCR5 deletion on the growth of *Yersinia pseudotuberculosis* and survival of mice after infection with *Y. pestis*. **a**, Bacterial growth in the caecum or in Peyer's patches (PP) of mice infected by orogastric lavage (D. Monack) with 2×10^9 *Y. pseudotuberculosis* YPIIIIV. Bacterial colony-forming units (CFU) are per g tissue at 4 days after infection. Plot shows the median, the range from 25th to 75th percentile (boxed) and the data range. WT, normal wild-type C57BL/6 mice; R5-KO, CCR5-deficient C57BL/6 mice. **b**, Comparison of survival of BALB/c mice with and without CCR5 ($n = 40$ per group), over 10 days following intravenous challenge with 10^2 *Y. pestis* KIM, substrain D27 (Pgm – , LcrV+). Green, CCR5 WT females; orange and black, CCR5-deficient males and females, respectively. **c**, Survival of mice after intravenous challenge with about ten *Y. pestis* organisms, which is close to the LD₅₀ dose in this strain. Green, CCR5 WT BALB/c females; purple, CCR5 WT BALB/c males ($n = 20$ per group); red, CCR5 WT C.B-17 SCID females ($n = 8$); blue, CCR5 WT C.B-17 SCID males ($n = 4$).

zygotes have a reduction in surface expression of CCR5 and slower progression from HIV infection to AIDS⁵. The lack of CCR5 at the cell surface has no obvious deleterious effect in humans. Mice with homozygous deletion of CCR5 have subtle immunological abnormalities, including in controlling infection by intracellular organisms such as *Listeria*, *Cryptococci* and *Leishmania*^{7–9}.

To test whether CCR5 deficiency protects mice against *Yersinia* infection, we challenged them with lethal inocula of wild-type *Y. pseudotuberculosis* YPIIIpYV (C57BL/6 CCR5-deficient mice) or Pgm – *Y. pestis* KIM substrain D27 (BALB/c CCR5-deficient mice). There was no significant difference in the bacterial load in the caecum or Peyer's patches at two or four days post-infection between C57BL/6 CCR5-deficient and CCR5-expressing mice following oral infection (Fig. 1a). Macrophages from CCR5-deficient animals showed little to no difference in bacterial growth of *Y. pseudotuberculosis* or *Y. pestis* compared with those from CCR5-expressing mice.

These results argue against CCR5 being essential for infection by *Y. pestis* or *Y. pseudotuberculosis*. However, they do not eliminate the possibility that a protective effect caused by CCR5 deletion may reduce mortality without changing bacterial spread. We therefore evaluated the effect of CCR5 deficiency on survival after *Y. pestis* infection in the more susceptible BALB/c mouse strain, but found no significant difference in survival between CCR5-deficient BALB/c female mice and BALB/c females with normal CCR5 expression (Fig. 1b).

Male CCR5-deficient mice survived for a shorter time (Fig. 1b; $P < 0.0001$). We therefore challenged male and female BALB/c mice with a lower dose of *Y. pestis* and again found that the males showed significantly reduced survival (Fig. 1c; $P = 0.0006$). As increased susceptibility was also seen in limited numbers of male SCID mice (Fig. 1c),

this gender-related defect could be affected by innate immunity. The poorer survival of CCR5-deficient male mice (Fig. 1b) is therefore likely to be related to their gender and not to CCR5 deficiency. Gender may also be a factor in *Yersinia* infection in humans, because males seem to be more susceptible to bubonic plague than females¹⁰.

Our results show that CCR5 deficiency in mice does not protect against infection or death caused by experimental *Yersinia* infection, making it unlikely that the CCR5Δ32 allele protects against plague. A modelling study¹¹ reaches a similar conclusion, with smallpox instead of plague being proposed as the disease that selected for the CCR5Δ32 allele.

Joan Mecsas*†, Greg Franklin‡, William A. Kuziel†§, Robert R. Brubaker¶, Stanley Falkow*, Donald E. Mosier‡

*Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford, California 94305-5402, USA

‡Department of Immunology, The Scripps Research Institute, La Jolla, California 92037, USA

e-mail: dmosier@scripps.edu

§Section of Molecular Genetics and Microbiology and Institute of Cellular and Molecular Biology, University of Texas at Austin, Texas 78712, USA

¶Department of Microbiology, Michigan State University, East Lansing, Michigan 48824-1101, USA

†Present addresses: Department of Microbiology and Molecular Biology, Tufts University, Boston, Massachusetts 02111, USA (J.M.); Protein Design Labs, Fremont, California 94555, USA (W.A.K.).

1. Stephens, J. C. *et al. Am. J. Hum. Genet.* **62**, 1507–1515 (1998).

2. Libert, F. *et al. Hum. Mol. Genet.* **7**, 399–406 (1998).

3. Perry, R. D. & Fetherston, J. D. *Clin. Microbiol. Rev.* **10**, 35–66 (1997).

4. Liu, R. *et al. Cell* **86**, 367–377 (1996).

5. Dean, M. *et al. Science* **273**, 1856–1862 (1996).

6. Huang, Y. *et al. Nature Med.* **2**, 1240–1243 (1996).

7. Zhou, Y. *et al. J. Immunol.* **160**, 4018–4025 (1998).

8. Huffnagle, G. B. *et al. J. Immunol.* **163**, 4642–4646 (1999).

9. Sato, N. *et al. J. Immunol.* **163**, 5519–5525 (1999).

10. Butler, T. *Trans. R. Soc. Trop. Med. Hyg.* **83**, 458–460 (1989).

11. Galvani, A. P. & Slatkin, M. *Proc. Natl Acad. Sci. USA* **100**, 15276–15279 (2003).

Competing financial interests: declared none.

Reversible redox energy coupling in electron transfer chains

Artur Osyczka^{1*}, Christopher C. Moser¹, Fevzi Daldal² & P. Leslie Dutton¹

¹The Johnson Research Foundation, Department of Biochemistry and Biophysics, and ²Department of Biology, Plant Science Institute, University of Pennsylvania, Philadelphia, Pennsylvania, 19104, USA

* Permanent address: Jagiellonian University, Faculty of Biotechnology, Kraków, Poland

Reversibility is a common theme in respiratory and photosynthetic systems that couple electron transfer with a transmembrane proton gradient driving ATP production. This includes the intensely studied cytochrome *bc*₁, which catalyses electron transfer between quinone and cytochrome *c*. To understand how efficient reversible energy coupling works, here we have progressively inactivated individual cofactors comprising cytochrome *bc*₁. We have resolved millisecond reversibility in all electron-tunnelling steps and coupled proton exchanges, including charge-separating hydroquinone–quinone catalysis at the Q_o site, which shows that redox equilibria are relevant on a catalytic timescale. Such rapid reversibility renders popular models based on a semiquinone in Q_o site catalysis prone to short-circuit failure. Two mechanisms allow reversible function and safely relegate short-circuits to long-distance electron tunnelling on a timescale of seconds: conformational gating of semiquinone for both forward and reverse electron transfer, or concerted two-electron quinone redox chemistry that avoids the semiquinone intermediate altogether.

The challenge of engineering efficient photosynthetic and respiratory energy conversion is to favour productive electron and proton transfer reactions that generate or use membrane proton motive force ($\Delta\mu\text{H}^+$), while suppressing energy-wasting short-circuit reactions. Photosynthetic reaction centres favour productive light-induced charge separation and avoid wasteful charge-recombining short-circuits by expending much of the absorbed light energy to drive the forward charge-separating steps, and thereby slow the reverse uphill charge returns that make short-circuits more likely.

For other crucial membrane energy-coupled oxidoreductases, such as the cytochrome *bc*₁ and *b₆f* family, the modest driving force provided by substrates makes this strategy impossible. Furthermore, as $\Delta\mu\text{H}^+$ builds up, the net reaction can be reversed, as shown in mitochondrial cytochrome *bc*₁ and complex I by classic experiments that artificially added ATP to increase $\Delta\mu\text{H}^+$ and to stimulate “reverse electron flow”^{1,2} on a timescale of minutes. Some lithotrophic organisms apparently rely on the reverse electron flow through cytochrome *bc*₁ for growth^{3,4}. Despite these observations, contemporary models for energy conversion in cytochrome *bc*₁ (refs 5–16) neglect reverse reactions and the implications of reversibility on short-circuit vulnerability. These models fail if reversibility on a rapid catalytic timescale is fully proven.

Here we have used photosynthetic membranes of the bacterium *Rhodobacter capsulatus* to investigate cytochrome *bc*₁ reversibility. *R. capsulatus* provides the kinetic means for rapid, light-activated delivery of the substrate’s hydroquinone (QH₂) and oxidizes cytochrome *c* to cytochrome *bc*₁ (Fig. 1a), even as the cofactors are progressively knocked out genetically. This strategy avoids long-standing difficulties in resolving concurrent reactions in the ‘b-chain’ (comprising haem *b_L*, haem *b_H* and the quinone of the Q_i site) and the ‘c-chain’ (comprising the iron-sulphur centre FeS, haem *c*₁ and cytochrome *c* (mostly *c*₂ but also *c_y*)). This strategy also establishes a rigorous single-turnover activation of cytochrome *bc*₁ to oxidize and reduce only one quinone molecule at the Q_o site, where the Q pool and the b- and c-chains meet and where energy conversion is catalysed.

Drawing on substrate and cofactor redox midpoint potentials and their pH dependencies (Fig. 1b), we have exposed selected single-turnover cofactor knockout systems to a range of driving forces from exergonic to endergonic (Fig. 1c) to define, step by step,

the thermodynamic parameters and to reveal the timescale of reversibility of the operating cytochrome *bc*₁.

Electron transfer in cofactor knockouts

Figure 2 presents the flash-activated haem *b* reduction kinetics in cytochrome *bc*₁ with different combinations of cofactor knockouts. The system is initially poised at high redox potentials to oxidize the Q pool. The uninhibited wild-type system (Fig. 2a, black) monitors flash-generated QH₂ arriving at the Q_o site and the 8-ms concurrent reduction of both FeS and haem *b_L*, followed by rapid haem *b_L* to *b_H* electron transfer across the membrane. Subsequent reduction of Q_i to semiquinone by haem *b_H* completes the electrical charging of the membrane (Fig. 1a).

After reduction, FeS normally undergoes constrained diffusion to transfer the electron to haem *c*₁ (refs 17–19). In turn, haem *c*₁ reduces cytochrome *c* (data not shown). Antimycin eliminates Q_i function (Fig. 2a, green), so that the electron advances only as far as haem *b_H* to reveal its full reduced extent. There is no noticeable effect of inactivation of the Q_i site on the rate of Q_o site turnover and haem *b_H* reduction. Further addition of myxothiazol (Fig. 2a, red) inhibits Q_oH₂ oxidation in the first place and no haem *b* reduction is observed; this rules out any other routes to haem *b* reduction in these experiments.

In the first cofactor knockout (Fig. 2b), mutation of the methionine ligand of haem *c*₁ to leucine²⁰ markedly drops the redox midpoint potential by hundreds of millivolts and disables electron transfer from either cytochrome *c* or FeS. But oxidation of QH₂ at the Q_o site proceeds to the same extent and with kinetics essentially identical to the wild type, and reduction of FeS by Q_oH₂ is unimpeded. Once FeS is reduced, oxidation of a second Q_oH₂ is impossible and, unlike the wild-type cytochrome *bc*₁, this knockout is a *de facto* Q_o site single-turnover enzyme.

In the second cofactor knockout (Fig. 2c), insertion of two alanines in the hinge region between the FeS head group and the transmembrane anchor severely interferes with the normal movement of FeS, thereby locking it in the Q_o site position¹⁷. This also prevents communication between the Q_o site and haem *c*₁ and cytochrome *c*, and again provides rigorous single-turnover action. But like the *c*₁ knockout, the FeS-motion knockout yields the same kinetics and extents as the wild type. Clearly, the diffusive, large-

scale domain motion of FeS is not needed for fully competent Q_o site activity^{17,21}. The third knockout (Fig. 2d) replaces the haem b_H histidine ligand with an asparagine, and haem b_H is lost²², providing an alternative way to restrict the Q_o site to one turnover. But reduction of haem b_L , together with reduction of FeS and haem c_1 , takes place with kinetics similar to that of haem b_H reduction.

The simultaneous knockout of two cofactors most severely strains cytochrome bc_1 turnover and delineates the thermodynamic limits of Q_o site action under high redox potential conditions. Pairing the b_H knockout with either the c_1 knockout (Fig. 2e) or the FeS-motion knockout (Fig. 2f) trims the multifactor enzyme to just three components: Q_o , FeS and haem b_L . Remarkably, Q_o site action has essentially the same rate of haem b reduction as the wild type. However, the extent of haem b reduction is only about a third under these conditions, indicating that the oxidation–reduction reaction of this simple three-component system may be energetically balanced and fully reversible on this timescale.

The reversibility of the system and the thermodynamically cooperative behaviour of the b- and c-chains become obvious when, following the guidelines of Fig. 1b and c, we manipulate the driving force provided to cytochrome bc_1 by changing the pH and state of the Q pool reduction before the activation. Figure 3 shows the pH modulation of the extent of haem b reduction for wild type and the c_1 knockout (Fig. 3a), and for the b_H knockout and the $b_H c_1$ double knockout (Fig. 3b) at both the high redox potential condition of the oxidized Q pool described above, and at a lower redox potential at which the Q pool is half-reduced and the arrival of QH_2 at the Q_o site is not rate-limiting. All systems become

progressively less competent as the driving force is lessened, either through raising the redox poise or through lowering the pH, which raises the midpoint potential of the Q pool (Fig. 1b, c).

A simple equilibrium model

The progressive failure of the extent of haem b reduction is neatly accommodated in an equilibrium model with four simple postulates (Fig. 3a, b, lines, and Fig. 3c–e, graphic illustrations). First, individual redox centres have the same midpoint potentials on the catalytic timescale as those measured in equilibrium redox titrations on a timescale of minutes (Fig. 1b). Second, the 2:1 ratio of reaction centres to cytochrome bc_1 monomer produces two oxidized cytochromes c_2 and one QH_2 per flash (ref. 9 and Fig. 3d); this postulate may be less valid at the high pH limit of 10, where QH_2 production by the Q_B site in reaction centres may begin to fail. Third, Q_o site redox reactions are strictly coupled so that every electron exchanged between the Q_o site quinone and FeS is accompanied by electron exchange between quinone and haem b_L (refs 23, 24). Fourth, electron transfers between Q_o and FeS and haem b_L , as well as electron transfers between members of the b-chain and between members of the c-chain, continue until the redox potential of the Q pool equals the average of the redox potentials of the c- and b-chain; that is, the net driving force for the Q_o site reaction is zero (refs 23, 24, and Fig. 3e).

The double knockout system (Fig. 3b, black) has the least driving force and is the first to fail as the pH is lowered (see Fig. 1c). When communication with haems c_1 and c_2 is restored in the single b_H knockout (Fig. 3b), the thermodynamically cooperative involvement of these haems improves cytochrome bc_1 robustness. At high redox potential (Fig. 3b, red), simple, rapid redox equilibrium contact with multiple oxidized redox centres in the c-chain favours more QH_2 oxidation and haem b_L reduction, which continues even

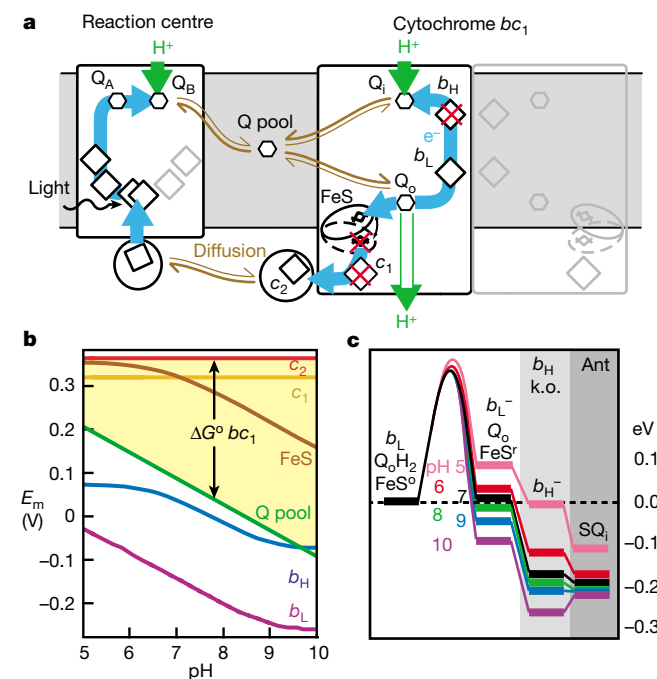


Figure 1 Cofactors and their energetics in *R. capsulatus*. **a**, Light-activated *R. capsulatus* initiates electron (blue) and proton (green) transfers through haems and chlorins (squares), quinones (hexagons) and the FeS cluster (double cross) of its reaction centre and cytochrome bc_1 to generate $\Delta\mu H^+$. The reaction centre generates oxidized cytochrome c_2 and hydroquinone (QH_2), which diffuse to a dimeric cytochrome bc_1 . Oxidized cytochrome c_2 oxidizes haem c_1 . Haem c_1 oxidizes FeS, which, by limited diffusion, arrives at the Q_o site to oxidize QH_2 drawn from the pool. Q_oH_2 oxidation reduces both FeS and haem b_L . Red crosses indicate the positions of cofactor knockouts. **b, c**, pH dependence of cofactor equilibrium midpoint (E_m) potentials (**b**) yields adjustable patterns of endergonic and exergonic steps beginning at the Q_o site (**c**). Grey areas indicate steps inactivated by antimycin and the b_H knockout.

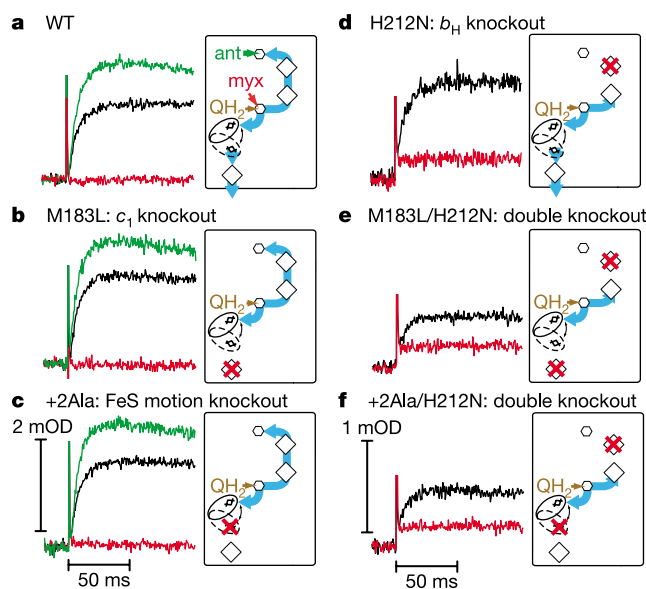


Figure 2 Flash-activated haem b reduction. Kinetics are shown for wild type (**a**) and cofactor knockout combinations (**b–f**) at pH 9.0 and at an E_h of +250 mV to oxidize the Q pool, FeS and haem b before the flash. Haem kinetics, initiated by QH_2 diffusing from the reaction centre, is uninhibited (black) or inhibited by antimycin (green), antimycin and myxothiazol (red, left), or myxothiazol alone (red, right). Reduction of haem b_H (left) and haem b_L (right) is monitored by absorption change presented in milli-units of optical density (mOD) at 560–570 nm and 566–573 nm, respectively; the initial step reflects reaction centre spectral contributions at these wavelengths. Red crosses indicate the locations of knockouts and blue arrows delineate remaining electron transfer reactions.

as the Q pool midpoint rises with lower pH, until about pH 7. At redox potentials with the Q pool half-reduced (Fig. 3b, green), more driving force is provided and failure is delayed until the pH is lower than 6. Similarly, when communication with haem b_H is restored and the contribution of haems c_1 and c_2 is denied in the c_1 knockout (Fig. 3a, grey), the thermodynamic assistance of haem b_H cooperates in maintaining haem b reduction until pH 6. In wild type, both at high (Fig. 3a, blue) and low (Fig. 3a, magenta) redox potentials, the cooperative action of the extended c- and b-chains allows haem b_H reduction and transmembrane electron transfer to continue well below pH 6. The simple equilibrium model tracks the progressive failure of all of the different systems.

The thermodynamic cooperativity of the b-chain apparent in the b_H knockouts is likely to be enhanced by cross-dimer electron transfer (Fig. 4). Cytochrome bc_1 structures^{25–28} reveal a tunnelling distance between the two haems b_L in the dimer that is only slightly longer than the tunnelling distance between haem b_L and haem b_H ; conservative calculation places haem b_L to b_L electron tunnelling at tens of microseconds²⁹. This means that whenever Q_o site catalysis in one or both monomers leaves two electrons in the b-chain of the dimeric complex, tunnelling between components of the chain should allow both electrons to doubly reduce a single Q_i on a millisecond timescale.

Significantly, this redistribution of electrons can occur without involving the Q_o site. In effect, this removes the strict coupling between two turnovers of one Q_o site and one Q_i site described in

the traditional double Q-cycle model³⁰. It also helps to explain how the first substoichiometric fraction of antimycin that binds, inhibits noticeably less effectively than the final fraction^{31,32}.

Although the extent of haem b reduction is considerably modulated by changing the pH and by the addition and removal of redox centres, the reaction rate is unperturbed for the most part (Fig. 5). At high redox potentials when the Q pool is initially oxidized (Fig. 5, blue), pH has little effect on the QH_2 oxidation rate; under these conditions, the process is simply limited by the diffusion of QH_2 from the reaction centre to cytochrome bc_1 . At low redox potentials (Fig. 5, red), where QH_2 is already available in the pool and not rate-limiting, lowering the pH below 7 finally begins to slow haem b reduction (Fig. 5, dotted line limit is 10-fold change of rate per pH unit). This drop in rate parallels the shrinking driving force for QH_2 to FeS and haem b_L electron transfer, as determined by the equilibrium midpoint potentials (Fig. 1b).

The b- and c-chain components contribute directly and independently to the total driving force of the reaction. The *de facto* single-turnover knockouts show that the haem b and haem c_1 re-reduction rates are not dependent on haem b_L to b_H electron transfer and Q_i site action in the b-chain, as was previously suggested^{14,16,33}. Independence is similarly clear when exogenous light-activated oxidants trigger a single turnover showing no antimycin effect on the haem c_1 re-reduction rates³⁴. The knockouts show no indication of the proposed long-range effects of Q_i on electron transfer in the c-chain³⁵, of tight coupling of FeS motion to

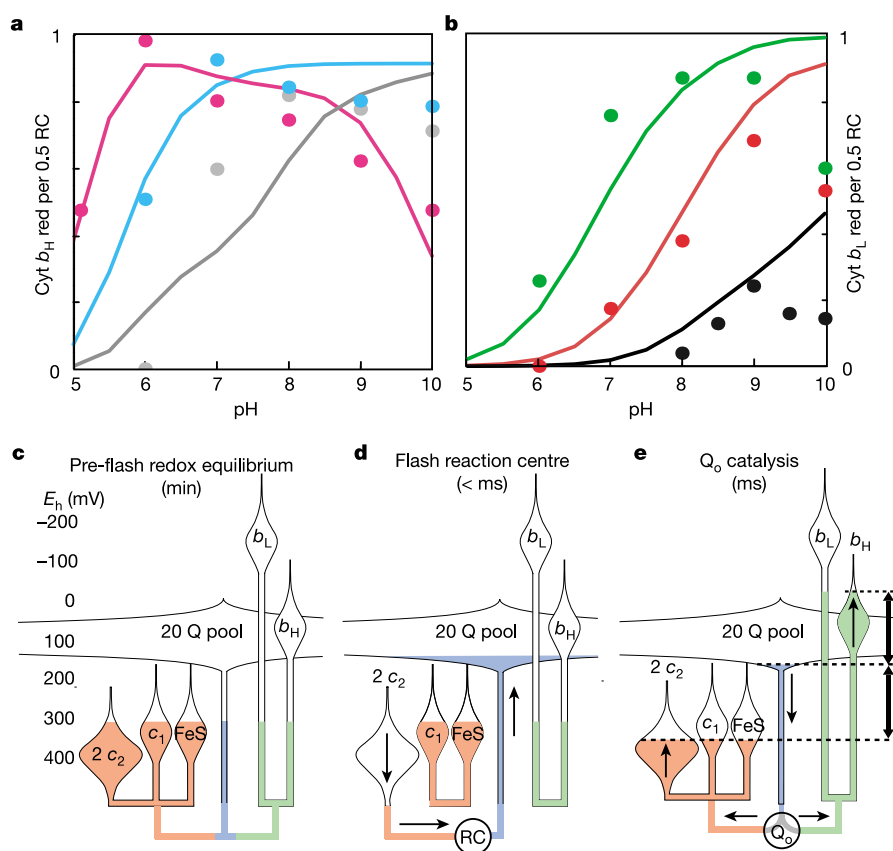


Figure 3 Extent of flash-activated Q_o site catalysis fitted to the simple equilibrium model. Points indicate yields of haem b reduction catalysed by the Q_o site, lines are derived from the equilibrium model. **a**, Haem b reduction yields (antimycin) in wild type with the Q pool half-reduced (magenta) or oxidized (blue), and in the c_1 knockout with the Q pool oxidized (grey). **b**, Haem b reduction yields in the b_H knockout with the Q pool half-reduced (green) or oxidized (red), and in the $c_1 b_H$ double knockout with the Q pool oxidized (black).

c, Simple equilibrium model: redox equilibration on a timescale of minutes (E_h) sets initial equilibrium levels of reduction of the c-chain (red), Q pool (blue) and b-chain (green). **d**, Flash-activated reaction centre (RC) rapidly offsets reduction levels of cytochrome c_2 and Q/QH_2 . **e**, Millisecond reversible Q_o site catalysis continues until the Q pool potential is the average of the c- and b-chains. Flared reservoirs reflect nernstian redox buffering.

electron transfer in the b-chain^{8,14}, nor of regulation of electron transfer competence by carefully controlled protonation states^{5,7,36}.

Because a simple equilibrium model sufficiently explains reaction extents under an extreme range of thermodynamic and knockout conditions, there is little reason to suppose that redox properties on the millisecond timescale are substantially different from the equilibrium values. There is no evidence that redox states of one centre profoundly influence the redox properties of others, and the influence of protonation or deprotonation of a redox centre on its affinity for electrons is adequately reflected in the equilibrium pH dependencies of these centres.

Short-circuiting of reversible electron transfer

The short-circuit rate of haem *b* reoxidation when Q_i is inoperative (Fig. 5, purple) shows that the rapidly reversible cytochrome bc_1 , like the essentially irreversible reaction centre, suppresses short-circuits into the timescale of seconds, even though the methods of suppression must be different. Figure 6 examines structural contributions to physiologically reversible cytochrome bc_1 and attendant short-circuits. The two short-circuit reactions of Fig. 6a and b involve oxidation of QH_2 by FeS and haem c_1 , or reduction of Q by haem b_L and haem b_H , and compete directly with the forward and reverse physiological reactions. However, both these short-circuits are prevented simply by the electron tunnelling geometry of cytochrome bc_1 . Haem b_L and FeS are positioned close to Q_o , fostering submillisecond physiological tunnelling, while the potentially short-circuiting haem b_H and haem c_1 components are considerably farther away, with corresponding tunnelling rates slowed safely to a timescale of seconds. A natural consequence of the distant haem c_1 is that FeS must undergo motion between the haem c_1 and the Q_o site^{17–19} to achieve short enough tunnelling distances to keep electron transfer faster than the catalytic rate. Although rapid FeS motion may have a role in reassembling and rapidly refreshing the Q_o site with protons and quinones to reflect the equilibrium pH and quinone pool redox states, in itself the simple motion does not prevent the short-circuit reactions of Fig. 6.

The three-step short-circuits of Fig. 6c and d might be minimized with sufficiently unstable semiquinone (that is, with a small stability constant, K_{stab})^{9,24,37,38}. Simple tunnelling calculations²⁹, using crystal structure distances^{23–28} and driving force (ΔG°) values given by

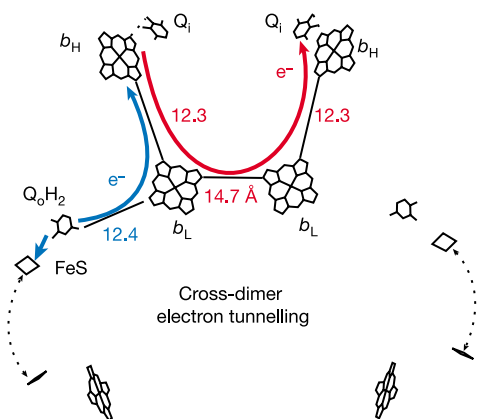


Figure 4 Intermonomer tunnelling in cytochrome bc_1 . After initial entry of an electron into a b-chain (blue arrow), distances between b_L haems in the dimer are short enough to permit tunnelling equilibration between all *b* haems and Q_i sites (red arrow) on a timescale of Q_o site catalysis and in a manner independent of the Q_o site.

equilibrium titrations (Fig. 1b) and a range of K_{stab} values for semiquinone, indicate how unstable the semiquinone must be to suppress short-circuits. With quinone in a position similar to the inhibitor stigmatellin and a reorganization energy for Q_o electron transfer similar to that of Q_B (1.0–1.2 eV), a K_{stab} value of less than 10^{-16} is required to slow the three-step short-circuit reactions of Fig. 6c and d to a timescale of seconds. Even smaller values ($K_{stab} < 10^{-22}$) are required to slow the two-step short-circuits of Fig. 6e and f.

A low K_{stab} slows short-circuits by making some reactions, such as haem b_L to oxidized Q_o electron transfer, extremely uphill; however, the physiologically productive forward and reverse reactions that use these same uphill steps will also be slowed. With the geometry and energetics described above, K_{stab} must be greater than 10^{-10} to allow haem b_L to Q_o electron tunnelling on a submillisecond timescale in reverse physiological electron transfer. Thus, there is no single value of K_{stab} that will permit rapid submillisecond forward and reverse electron tunnelling and simultaneously limit short-circuit tunnelling to the timescale of seconds.

Safe, reversible Q_o site energy conversion

Nearly all proposed models^{5–8,10–12,14,16} invoke redox-sensitive conformational choreography to gate the stability of the semiquinone state (see Supplementary information), and all fail because they neglect or even forbid reverse reactions and do not consider oxidized Q_o -mediated short-circuits. For example, the model of a catalytic switch of FeS⁸ supposes that FeS cannot assume the 'b' position at the Q_o site and react with Q_o unless both FeS and haem b_L are oxidized; this avoids the short-circuits shown in Fig. 6c, e. However, this model prevents neither the short-circuit of Fig. 6f, because when haem b_L reduces Q_o , both FeS and haem b_L are oxidized and FeS will be reduced by highly favourable electron transfer from semiquinone, nor the short-circuit of Fig. 6d, in which successive FeS-independent electron transfers from haem b_L reduce Q_o . Furthermore, the reverse physiological reaction, in which haem b_L and then FeS reduce Q_o to Q_oH_2 , seems to be forbidden because

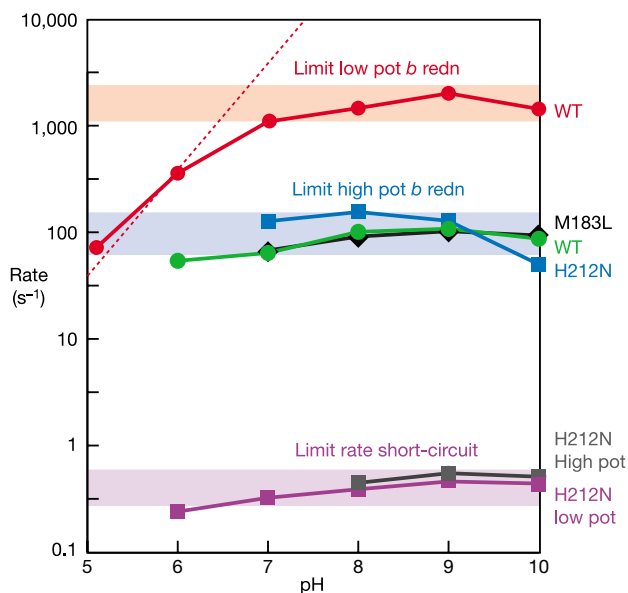


Figure 5 Rates of Q_i -inhibited haem *b* reduction and reoxidation in cofactor knockouts as a function of pH. When the Q pool is half-reduced before the flash (red), rates follow the limiting catalytic rate, slowing at the lowest pHs where the driving force for Q_o site catalysis decreases (Fig. 1c). When the Q pool is oxidized before the flash (blue), rates follow the rate of diffusion of QH_2 from the reaction centre. Haem *b* reoxidation rates (purple) reflect energy-wasting short-circuits; these are unaffected by added redox-poising dyes.

the reduced FeS is assumed not to take the 'b' position.

Any attempt to repair these models requires additional redox-sensitive conformational gating and explicit reversibility. This is most easily achieved by double gating of the Q_o site (see Supplementary Information). Double gating could be based on controlled hydroquinone and quinone binding or on modulated semiquinone redox properties such as K_{stab} to enable semiquinone formation when FeS and haem b_L are both oxidized (physiological forward) or both reduced (physiological reverse) and not when FeS is oxidized and haem b_L reduced. Application of such gating elements successfully allows reversibility while eliminating the short-circuits of Fig. 6c–f.

An alternative abandons the elaborations implicit in gated sequential models and exploits the potential for reversible, genuinely 'concerted' electron transfer provided by the quinone redox chemistry itself. We use concerted as a 'kinetic' term that defines two events taking place without time for significant atomic rearrangement, in other words, within femtoseconds³⁹. Theoretical and experimental descriptions of kinetically concerted two-electron transfers are developing^{40,41}. In contrast to common sequential models (Fig. 6h, bottom), this fundamentally different mechanism for Q_o site catalysis has reversible concerted two-electron transfer taking place through a transition state devoid of semiquinone character (Fig. 6h, top).

The term concerted has long been used in the cytochrome bc_1 literature^{12,24,42}, but it usually refers to the 'thermodynamic' cooperativity of the Q_o site reaction, in which electron transfer is sequential with a transient semiquinone state that never accumulates enough to be observable; in this case, reduction and oxidation of reaction partners may appear concurrent. The term concerted continues to be used in redox-sensitive gating models¹⁵, even though kinetic concertedness renders such gating meaningless.

A Q_o site designed for concerted catalysis must strongly destabi-

lize the semiquinone to a K_{stab} value of less than 10^{-22} to make sequential short-circuits through semiquinone energetically unfavourable enough to slow them to seconds. At the same time, the availability of proton donors and acceptors is essential^{43,44}. It is the combination of these circumstances that can favour concerted two-electron transfer reactions over sequential reactions^{39,45}. Although a concerted mechanism will probably have a larger contribution to the kinetic barrier from increased reorganization energy, this may be more than compensated by the markedly smaller ΔG° contribution to the barrier ($\Delta G^\circ_{concerted} = \text{near zero} \ll \Delta G^\circ_{sequential}$). Most of the thermal energy required to reach a concerted Q_o transition state presumably would be invested in reorganization in the form of occasionally synchronous movement of protons towards or away from the quinone oxygens, with the corresponding changes in bond lengths in the quinone.

The concerted model is silent on the exact sequence of protonation events relative to the two near simultaneous electron transfers, although it is likely that reorganization of the Q_o site moves at least one proton into an intermediate, non-equilibrium position to achieve the lowest energy transition state for rapid two-electron transfer. In the reversible concerted model, the only short-circuit reaction that remains (Fig. 6g) is over the much longer distance of 23 Å directly between haem b_L and FeS, which, like the same distance–charge recombination in reaction centres, should have a tunnelling time on the scale of seconds²⁹. Indeed, the short-circuit rate observed in cytochrome bc_1 (Fig. 5, purple) matches this predicted rate of seconds.

The models developed here show that efficient redox energy coupling in cytochrome bc_1 is assured by redox cofactor arrangement and different quinone chemistry at the Q_o and Q_i sites. Redox cofactors engaging in productive forward and reverse electron transfers are separated by tunnelling distances of less than about 14 Å. Within this limit, thermodynamically cooperative interactions

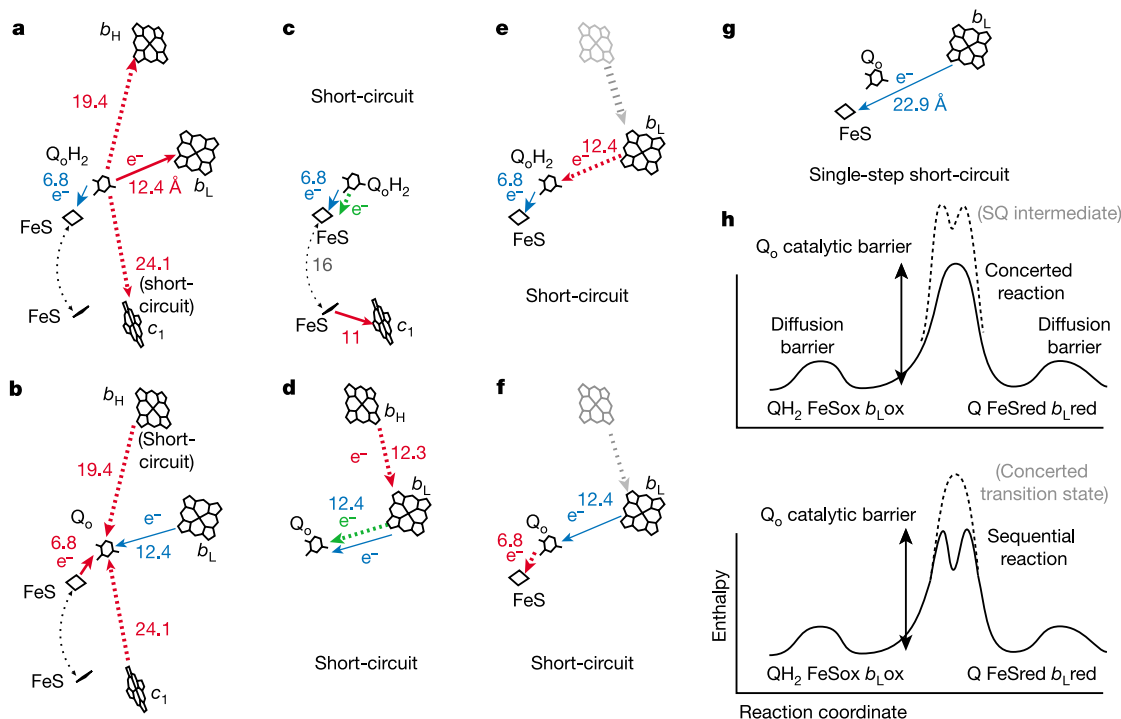


Figure 6 Q_o site catalysis and short-circuits in physiologically reversible cytochrome bc_1 . **a–f**, The geometry and spacing of cofactors^{25–28} is organized to promote forward and reverse physiological electron transfers (**a**, **b**, solid arrows) and to prevent some short-circuits (**a**, **b**, broken arrows) but not others (**c–f**). The order of electron transfers in models involving semiquinone intermediates is blue, red and then green. **g**, A short-circuit

that even a reversible concerted model cannot avoid involves long-distance and safely slow, direct electron tunnelling between haem b_L and FeS. **h**, Concerted (top) and sequential (bottom) mechanisms differ considerably in the reaction energy surface profile during forward and reverse Q_o site catalysis.

through the extended redox chains sustain mechanistically robust function over a broad range of redox and pH conditions. Unproductive electron transfers generally face longer distances. In the Q_o site, where such long-distance suppression of unproductive reactions is not possible because the same centres participate in productive electron transfer, reversible double gating of sequential reactions or a concerted ubiquinone redox chemistry is required to avoid semiquinone-mediated short-circuits. The Q_o site design to suppress or even exclude semiquinone joins the cross-dimer electron transfer action to sweep the b-chain of reduced haem *b*, diminishing the potential for the generation of damaging radicals and reactive oxygen species in normally functioning mitochondria⁴⁶. □

Methods

Cofactor knockouts

The single mutations in haem c_1 and the FeS subunit of *R. capsulatus* cytochrome bc_1 , M183L and the +2Ala insertion, have been described^{17,20}. The single H212N mutation in the cytochrome *b* subunit results in the selective loss of haem b_H and the retention of haem b_L in a manner similar to that reported in studies of *Rhodobacter sphaeroides* cytochrome bc_1 (ref. 22, and T. Dogerthy, K. Gray and F. Daldal, unpublished data). We created the double mutations M183L/H212N and +2Ala/H212N by ligating appropriate restriction fragments derived from plasmids carrying single mutations, and we introduced them into a suitable genetic background as described⁴⁷.

Membranes, titrations and kinetics

Chromatophore membrane preparation, dark equilibrium redox titrations of FeS and haem *b*, and light-induced time-resolved kinetics in the presence of valinomycin were done as described^{9,20,48,49}. The extent of haem *b* reduction was simulated with Mathematica (Wolfram Research) using the four postulates described in the text.

Received 11 August; accepted 14 November 2003; doi:10.1038/nature02242.

1. Chance, B. & Hollunger, G. The interaction of energy and electron transfer reactions in mitochondria. *J. Biol. Chem.* **236**, 1562–1568 (1961).
2. Miki, T., Miki, M. & Orii, Y. Membrane potential-linked reversed electron transfer in the beef heart cytochrome bc_1 complex reconstituted into potassium-loaded phospholipid vesicles. *J. Biol. Chem.* **269**, 1827–1833 (1994).
3. Elbehti, A., Brasseur, G. & Lemesle-Meunier, D. First evidence for existence of an uphill electron transfer through the bc_1 and NADH-Q oxidoreductase complexes of the acidophilic obligate chemolithotrophic ferrous ion-oxidizing bacterium *Thiobacillus ferrooxidans*. *J. Bacteriol.* **182**, 3602–3606 (2000).
4. Nicholls, D. G. & Ferguson, S. J. *Bioenergetics 3* (Academic Press, San Diego, 2002).
5. Link, T. A. The role of the 'Rieske' iron sulfur protein in the hydroquinone oxidation (Q_p) site of the cytochrome bc_1 complex. The 'proton-gated affinity change' mechanism. *FEBS Lett.* **412**, 257–264 (1997).
6. Brandt, U. & von Jagow, G. Analysis of inhibitor binding to the mitochondrial cytochrome-*c* reductase by fluorescence quench titration—evidence for a 'catalytic switch' at the Q_o center. *Eur. J. Biochem.* **195**, 163–170 (1991).
7. Brandt, U. Bifurcated ubihydroquinone oxidation in the cytochrome bc_1 complex by proton-gated charge transfer. *FEBS Lett.* **387**, 1–6 (1996).
8. Brandt, U. The chemistry and mechanics of ubihydroquinone oxidation at center P (Q_o) of the cytochrome bc_1 complex. *Biochim. Biophys. Acta* **1365**, 261–268 (1998).
9. Ding, H. *et al.* Ubiquinone pair in the Q_o site central to the primary energy conversion reactions of cytochrome bc_1 complex. *Biochemistry* **34**, 15979–15996 (1995).
10. Crofts, A. R. *et al.* Pathways for proton release during ubihydroquinone oxidation by the bc_1 complex. *Proc. Natl Acad. Sci. USA* **96**, 10021–10026 (1999).
11. Crofts, A. R. *et al.* Mechanism of ubiquinol oxidation by the bc_1 complex: different domains of the quinol binding pocket and their role in the mechanism and binding of inhibitors. *Biochemistry* **38**, 15807–15826 (1999).
12. Snyder, C. H., Gutierrez-Cirlos, E. B. & Trumpower, B. L. Evidence for a concerted mechanism of ubiquinol oxidation by the cytochrome bc_1 complex. *J. Biol. Chem.* **275**, 13535–13541 (2000).
13. Dutton, P. L. *et al.* In *Coenzyme Q: Molecular Mechanisms in Health and Disease* (eds Kagan, V. E. & Quinn, P. J.) 65–82 (CRC, Boca Raton, FL, 2000).
14. Yu, C. A., Wen, X. L., Xiao, K. H., Xia, D. & Yu, L. Inter- and intra-molecular electron transfer in the cytochrome bc_1 complex. *Biochim. Biophys. Acta* **1555**, 65–70 (2002).
15. Trumpower, B. L. A concerted, alternating sites mechanism of ubiquinol oxidation by the dimeric cytochrome bc_1 complex. *Biochim. Biophys. Acta* **1555**, 166–173 (2002).
16. Klishin, S. S., Junge, W. & Mulikjanian, A. Y. Flash-induced turnover of the cytochrome bc_1 complex in chromatophores of *Rhodobacter capsulatus*: binding of Zn^{2+} decelerates likewise the oxidation of cytochrome *b*, the reduction of cytochrome c_1 and the voltage generation. *Biochim. Biophys. Acta* **1553**, 177–182 (2002).
17. Darrouzet, E., Valkova-Valchanova, M. B., Moser, C. C., Dutton, P. L. & Daldal, F. Uncovering the [2Fe2S] domain movement in cytochrome bc_1 and its implications for energy conversion. *Proc. Natl Acad. Sci. USA* **97**, 4567–4572 (2000).
18. Xiao, K., Yu, L. & Yu, C. A. Confirmation of the involvement of protein domain movement during the catalytic cycle of the cytochrome bc_1 complex by the formation of an intersubunit disulfide bond between cytochrome *b* and the iron-sulfur protein. *J. Biol. Chem.* **275**, 38597–38604 (2000).
19. Darrouzet, E., Moser, C. C., Dutton, P. L. & Daldal, F. Large scale domain movement in cytochrome bc_1 : a new device for electron transfer in proteins. *Trends Biochem. Sci.* **26**, 445–451 (2001).

20. Gray, K. A., Davidson, E. & Daldal, F. Mutagenesis of methionine-183 drastically affects the physicochemical properties of cytochrome c_1 of the bc_1 complex of *Rhodobacter capsulatus*. *Biochemistry* **31**, 11864–11873 (1992).
21. Darrouzet, E. & Daldal, F. Movement of the iron-sulfur subunit beyond the *ef* loop of cytochrome *b* is required for multiple turnovers of the bc_1 complex but not for single turnover Q_o site catalysis. *J. Biol. Chem.* **277**, 3471–3476 (2002).
22. Yun, C.-H., Crofts, A. R. & Gennis, R. B. Assignment of the histidine axial ligands to the cytochrome b_H and cytochrome b_L components of the bc_1 complex from *Rhodobacter sphaeroides* by site-directed mutagenesis. *Biochemistry* **30**, 6747–6754 (1991).
23. Mitchell, P. The protonmotive Q cycle: a general formulation. *FEBS Lett.* **59**, 137–139 (1975).
24. Mitchell, P. Possible molecular mechanisms of the proton motive function of cytochrome systems. *J. Theor. Biol.* **62**, 327–367 (1976).
25. Xia, D. *et al.* Crystal structure of the cytochrome bc_1 complex from bovine heart mitochondria. *Science* **277**, 60–66 (1997).
26. Iwata, S. *et al.* Complete structure of the 11-subunit bovine mitochondrial cytochrome bc_1 complex. *Science* **281**, 64–71 (1998).
27. Zhang, Z. *et al.* Electron transfer by domain movement in cytochrome bc_1 . *Nature* **392**, 677–684 (1998).
28. Hunte, C., Koepke, J., Lange, C., Roßmanith, T. & Michel, H. Structure at 2.3 Å resolution of the cytochrome bc_1 complex from the yeast *Saccharomyces cerevisiae* co-crystallized with an antibody F_v fragment. *Structure* **8**, 669–684 (2000).
29. Page, C. C., Moser, C. C. & Dutton, P. L. Natural engineering principles of electron tunnelling in biological oxidation–reduction. *Nature* **402**, 47–52 (1999).
30. Crofts, A. R., Meinhardt, S. W., Jones, K. R. & Snozzi, M. The role of the quinone pool in the cyclic electron-transfer chain of *Rhodopseudomonas sphaeroides*: a modified Q-cycle mechanism. *Biochim. Biophys. Acta* **723**, 202–218 (1983).
31. Berden, J. A. & Slater, E. C. The allosteric binding of antimycin to cytochrome *b* in the mitochondrial membrane. *Biochim. Biophys. Acta* **256**, 199–215 (1971).
32. Gutierrez-Cirlos, E. B. & Trumpower, B. L. Inhibitory analogs of ubiquinol act anti-cooperatively on the yeast cytochrome bc_1 complex. *J. Biol. Chem.* **277**, 1195–1202 (2002).
33. Kim, H. *et al.* Inhibitor binding changes domain mobility in the iron-sulfur protein of the mitochondrial bc_1 complex from bovine heart. *Proc. Natl Acad. Sci. USA* **95**, 8026–8033 (1998).
34. Sadowski, R. *et al.* Use of a photoactivated ruthenium dimer complex to measure electron transfer between the Rieske iron-sulfur protein and cytochrome c_1 in the cytochrome bc_1 complex. *Biochemistry* **39**, 4231–4236 (2000).
35. Lange, C. & Hunte, C. Crystal structure of the yeast cytochrome bc_1 complex with its bound substrate cytochrome *c*. *Proc. Natl Acad. Sci. USA* **99**, 2800–2805 (2002).
36. Guergova-Kuras, M., Kuras, R., Ugulava, N., Hadad, N. & Crofts, A. R. Specific mutagenesis of the Rieske iron-sulfur protein in *Rhodobacter sphaeroides* shows that both the thermodynamic gradient and the pK of the oxidized form determine the rate of quinol oxidation by the bc_1 complex. *Biochemistry* **39**, 7436–7444 (2000).
37. Junemann, S., Heathcote, P. & Rich, P. R. On the mechanism of quinol oxidation in the bc_1 complex. *J. Biol. Chem.* **273**, 21603–21607 (1998).
38. Berry, E. A., Zhang, Z., Huang, L.-S. & Kim, S.-H. Structure of quinone-binding sites in *bc* complexes: functional implications. *Biochem. Soc. Trans.* **27**, 565–572 (1999).
39. Gileadi, E. Simultaneous two-electron transfer in electrode kinetics. *J. Electroanal. Chem.* **532**, 181–189 (2002).
40. Boman, P. *et al.* Kinetics of bond shift and charge transfer in dialkynylphenylene-bridged dicyclocatetraenes and their dianions. *J. Am. Chem. Soc.* **121**, 1558–1564 (1999).
41. Lambert, C. Concerted two-electron-transfer processes in mixed-valence species with square topology. *Chemphyschem* **4**, 877–880 (2003).
42. Kramer, D. M. & Crofts, A. R. The concerted reduction of the high- and low-potential chains of the *bf* complex by plastoquinol. *Biochim. Biophys. Acta* **1183**, 72–84 (1993).
43. Rich, P. R. Electron and proton transfers in chemical and biological quinone systems. *Discuss. Faraday Soc.* **74**, 349–364 (1982).
44. Greaves, M. D., Niemz, A. & Rotello, V. M. Control of one- versus two-electron reduction of ubiquinone via redox-dependent recognition. *J. Am. Chem. Soc.* **121**, 266–267 (1999).
45. Zusman, L. D. & Beratan, D. N. Three-state model for two-electron transfer reactions. *J. Phys. Chem. A* **101**, 4136–4141 (1997).
46. St-Pierre, J., Buckingham, J. A., Roebuck, S. J. & Brand, M. D. Topology of superoxide production from different sites in the mitochondrial electron transport chain. *J. Biol. Chem.* **277**, 44784–44790 (2002).
47. Atta-Asafo-Adjei, E. & Daldal, F. Size of the amino acid side chain at position 158 of cytochrome *b* is critical for an active cytochrome bc_1 complex and for photosynthetic growth of *Rhodobacter capsulatus*. *Proc. Natl Acad. Sci. USA* **88**, 492–496 (1991).
48. Osyczka, A., Dutton, P. L., Moser, C. C., Darrouzet, E. & Daldal, F. Controlling the functionality of cytochrome c_1 redox potentials in the *Rhodobacter capsulatus* cytochrome bc_1 complex through disulfide anchoring of a loop and a β -branched amino acid near the heme-ligating methionine. *Biochemistry* **40**, 14547–14556 (2001).
49. Robertson, D. E. *et al.* Discrete catalytic sites for quinone in the ubiquinol-cytochrome c_2 oxidoreductase of *Rhodopseudomonas capsulata*. *J. Biol. Chem.* **261**, 584–591 (1986).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements This work was supported by grants from the National Institutes of Health to P.L.D. and F.D.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to P.L.D. (dutton@mail.med.upenn.edu).

The central image of a gravitationally lensed quasar

Joshua N. Winn¹, David Rusin^{1,2} & Christopher S. Kochanek^{1,3}

¹Harvard-Smithsonian Center for Astrophysics, Cambridge, Massachusetts 02138, USA

²Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA

³Department of Astronomy, The Ohio State University, Columbus, Ohio 43210, USA

A galaxy can act as a gravitational lens, producing multiple images of a background object. Theory predicts that there should be an odd number of images produced by the lens^{1,2}, but hitherto almost all lensed objects have two or four images. The missing 'central' images, which should be faint and appear near the centre of the lensing galaxy, have long been sought as probes of galactic cores too distant to resolve with ordinary observations^{3–7}. There are five candidates for central images, but in one case the third image is not necessarily the central one^{8–10}, and in the others the putative central images might be foreground sources^{11–15}. Here we report a secure identification of a central image, based on radio observations of one of the candidates¹⁴. Lens models using the central image reveal that the massive black hole at the centre of the lensing galaxy has a mass of $< 2 \times 10^8$ solar masses ($M_\odot$), and the galaxy's surface density at the location of the central image is $> 20,000 M_\odot \text{pc}^{-2}$, which is in agreement with expectations based on observations of galaxies that are much closer to the Earth.

Central images are often called 'odd' images, but the key property that makes them interesting is not their odd number, but rather their proximity to the centre of the lens galaxy. To emphasize the distinction, we note that there can be an odd number of images without any central images, if the gravitational field has a large quadrupole⁹ or if there are multiple lens galaxies^{16,17}.

The absence of central images is a problem dating to the earliest days of galaxy lens observations, and many solutions have been proposed^{3,18,19}. Most recently, Evans and Hunter⁶ and Keeton⁷ argued that the absence of central images at current detection limits is no longer surprising, given recent Hubble Space Telescope observations of nearby galaxies²⁰. Those observations showed that the distribution of stars near the centres of massive elliptical galaxies (the predominant type of lens galaxy) is highly concentrated. Because the central image flux depends inversely on the square of the surface density, the concentrated density profiles should cause central images to be very faint (or even to have zero flux, if the density is singular), and detection limits will need to be improved by a factor of 10–50 before central images are commonly seen. In the meantime, the most favourable systems are radio-loud double quasars with large flux ratios. Radio waves are not extinguished by dust or overpowered by light in the lens galaxy, and asymmetric doubles produce the brightest possible central image for a given mass distribution.

One such system is the Parkes–MIT–NRAO (PMN) source J1632–0033 (ref. 14). It has two images (A and B) of a radio-loud quasar at redshift 3.42, with a flux ratio of 13 (Fig. 1). The lens redshift is unknown; our working assumption is $z_L = 1.0$, as estimated by requiring the galaxy's photometry and mass to be consistent with the fundamental plane of elliptical galaxies¹⁴. There is also a faint radio component (C) with a position and flux appropriate for a central image. However, as in a few other systems^{11–13}, the possibility could not be excluded that C is an active galactic nucleus in the lens galaxy rather than a third quasar image¹⁵. The simplest test is to compare the radio spectra of the components.

Because lensing preserves the frequency of photons, lensed images have the same spectrum (in the absence of differential propagation effects) whereas there is no reason why a foreground active galactic nucleus would have the same spectrum as a background source.

Previously, we attempted this comparison at frequencies from 1.7 to 15 GHz, finding that C was fainter at low frequency than expected for a third image¹⁵. However, the discrepancy was limited to a single measurement at the lowest frequency, where radio propagation effects are strongest. A central image might be affected more than other images by scintillation or absorption, owing to its passage through the dense galactic centre. Thus we could not come to a firm conclusion before obtaining data at higher frequencies, where propagation effects (scaling characteristically as ν^{-2}) are negligible.

We have now extended our measurements to 22 and 43 GHz, and obtained additional data at 8 and 15 GHz, using the Very Large Array. The high-frequency spectrum of the central component agrees well with those of the bright quasar images (Fig. 2). For $\nu > 1.7$ GHz, the logarithmic slopes of flux density ratio versus frequency (which should be zero for lensed images) are 0.00 ± 0.04 for B/A and -0.02 ± 0.07 for C/A. This is powerful evidence that C is a third quasar image.

The evidence that C is not only a third image, but also a long-sought central image, is its proximity to the centre of the lens galaxy (≈ 30 milli-arcsec, mas) and its faintness (0.41% the flux density of A). This sets PMN J1632–0033 apart from the only other three-image system known, Automatic Plate Measuring Survey (APM) 08279+5255, in which the lens galaxy has not been detected, and the fluxes of all three images are of the same order of magnitude. (Similar fluxes suggest that the third image in that system may not be a central image, but rather a 'naked-cusp' image due to a highly flattened mass distribution^{8,9}.)

For PMN J1632–0033, the central image properties can be used to constrain the core structure of the lens galaxy. In general, this requires detailed modelling, for which we refer the reader to ref. 15, in which some consequences of the central-image hypothesis were worked out before the status of C was clarified. Here we quote one result, and elaborate upon it to include the effect of a central black hole. If the mass distribution is taken to be a spherical power law, $\rho(r) \propto r^{-\beta}$, embedded in an external shear field to account for non-sphericity, then the data require $\beta = 1.91 \pm 0.02$ (2σ confidence), only slightly shallower than an isothermal ($\beta = 2$) mass distribution. The uncertainty is far smaller for this system than for typical two-image or four-image systems, because of the sensitive dependence of the central-image magnification on the exponent of the central cusp.

If the central mass were too large, then either a bright fourth image would be produced, or the central image would be demagnified out of existence^{3,6,7,21}. This fact can be used to derive an upper limit on the mass of any central black hole. We added a central

Table 1 Radio observations of PMN J1632–0033 with the Very Large Array

Date (UT)	Frequency (GHz)	A (mJy)	B (mJy)	C (mJy)	r.m.s. noise (mJy beam ⁻¹)
2003 Jul 23	8.46	202.13	16.49	0.84	0.033
2003 Aug 02	8.46	209.82	16.65	0.74	0.040
2003 Aug 30	8.46	195.19	15.86	0.78	0.042
2003 Jun 10	14.94	173.27	12.50	0.72	0.16
2003 Jul 01	14.94	180.17	13.87	0.57	0.13
2003 Jun 12	22.46	149.54	10.78	0.62	0.076
2003 Jun 24	22.46	146.95	10.81	0.48	0.061
2003 Jun 26	22.46	146.42	10.78	0.66	0.066
2003 Jun 15	43.34	107.25	7.85	0.38	0.077

The array was in the A configuration. The primary flux density calibrator was 3C286 and the phase calibrator was J1658+0741 (except the 8-GHz observation of 23 July 2003, for which it was J1651+0129). Initial calibration was performed with standard Astronomical Image Processing System (AIPS) procedures. Further self-calibration and imaging were performed with Difmap²². Flux densities were determined by fitting a model consisting of three point sources to the visibility data. For the 8-GHz and 15-GHz data, the relative positions of the components were fixed at the values determined by previous higher-resolution measurements¹⁴.

point mass to the power-law model described above, finding that the limits on β are barely affected and $M_{\text{BH}} < 2.0 \times 10^8 M_{\odot}$.

How does this compare to the mass of the black hole that is expected to reside in this galaxy? Among nearby galaxies, M_{BH} is correlated with the velocity dispersion of the surrounding stellar bulge^{22,23}. The exact correlation is a matter of debate, but for present purposes we adopt the correlation of ref. 24. Because the mass distribution of the lens galaxy in PMN J1632–0033 appears to be nearly isothermal, the A–B angular separation $\Delta\theta = 1.46$ arcsec can be used to estimate σ , using the relation $\Delta\theta/2 = (4\pi\sigma^2/c^2)(D_{\text{LS}}/D_{\text{S}})$, where D_{LS} and D_{S} are the lens–source and observer–source angular diameter distances. Assuming a flat $\Omega_{\text{M}} = 0.3$ cosmology with $H_0 = 72 \text{ km s}^{-1} \text{ Mpc}^{-1}$ and $z_{\text{L}} = 1.0$, the results are $\sigma = 224 \text{ km s}^{-1}$ and $M_{\text{BH}} = (2.1 \pm 0.4) \times 10^8 M_{\odot}$. Thus, the black hole cannot be much larger than expected from the local correlation.

A simple argument can also be used to place a lower bound on the core density of the galaxy. Because the density presumably does not increase with radius, the minimum central density occurs for the case of a constant-density core, for which the magnification of C is $\mu_{\text{C}} = (1 - \kappa_{\text{C}})^{-2}$. Here, κ is the surface density, in units of the cosmology- and redshift-dependent critical surface density for lensing. The near-isothermal models give $\mu_{\text{C}} = 0.0082$, from which it follows that $\kappa_{\text{C}} > 10$. Under the same cosmological assumptions as above, the surface density is $> 20,000 M_{\odot} \text{ pc}^{-2}$ within the radius of C. The radius of C is not known accurately but is $\lesssim 30 \text{ mas}$, corresponding to $\lesssim 230 \text{ pc}$, or $\lesssim 0.15 R_{\text{eff}}$ where R_{eff} is the effective radius of the galaxy.

How does this compare to nearby galaxy cores observed with the Hubble Space Telescope? Among the sample of ref. 20 are 38 early-type (E/S0) galaxies with measurements of core surface brightness profile, R_{eff} , and velocity dispersion. We used their estimates of the mass-to-light ratio (derived from the velocity dispersion) to convert the surface brightness profile into a surface density profile.

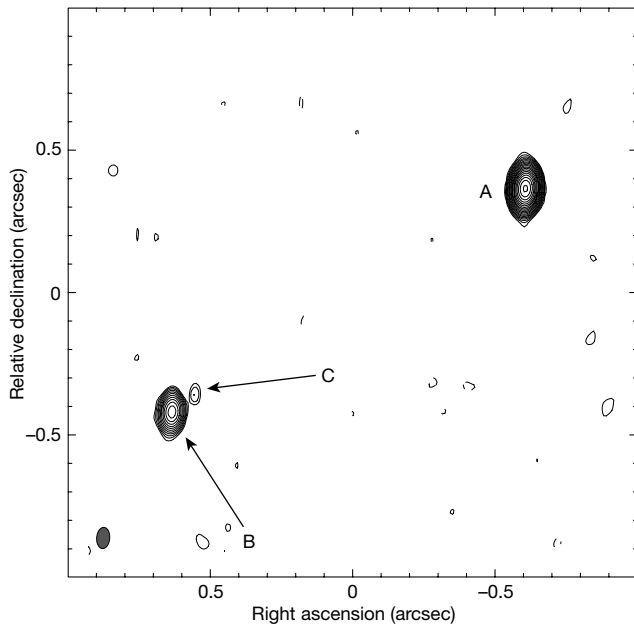


Figure 1 A radio map of the three components A, B and C of gravitational lens PMN J1632–0033. The observations were carried out with the Very Large Array at 43 GHz, with a total bandwidth of 100 MHz. Standard phase and amplitude calibration were applied, and the map was created with natural weighting. The root-mean-squared (r.m.s.) noise level in the residual map is $\sigma = 0.077 \text{ mJy beam}^{-1}$, as compared to the theoretical minimum r.m.s. of $0.04 \text{ mJy beam}^{-1}$. Contour levels begin at 2.5σ and increase by powers of $\sqrt{2}$. The synthesized beam (full-width at half-maximum of $0.073 \times 0.044 \text{ arcsec}$) is illustrated in the lower left corner of the map.

At $R/R_{\text{eff}} = 0.15$, the surface densities are $10^3\text{--}10^5 M_{\odot} \text{ pc}^{-2}$, showing that our result is consistent with local estimates and astrophysically relevant.

Thus we have derived limits on surface density and black-hole mass of the lens galaxy (co-moving distance ~ 3 gigaparsecs) that are consistent with more detailed measurements possible only for nearby galaxies ($\sim 30 \text{ Mpc}$). The lensing constraints have the advantage of depending directly on mass, rather than light. This motivates the discovery of additional central-image systems, and further observations of this system, some of which we describe below.

The most immediate challenges are direct measurement of the lens redshift and more accurate ($\lesssim 5 \text{ mas}$) measurement of the position of C relative to the lens centre. These data would better pin down the physical radius within which the limits on surface

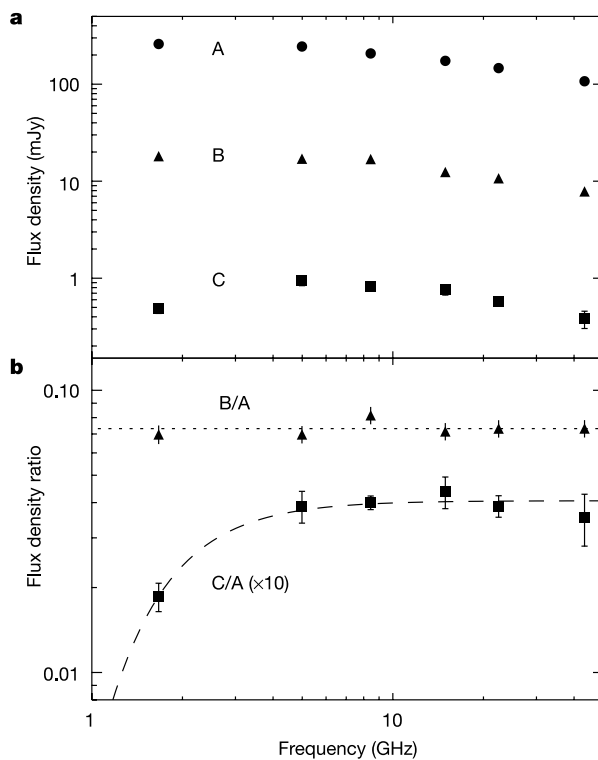


Figure 2 The central radio component and the bright quasar images have similar radio spectra. **a**, The flux density of each component as a function of frequency. **b**, The flux density ratios relative to A. Results for C/A were multiplied by ten for display purposes. All the observations at a given frequency from Table 1 and ref. 15 were combined by averaging the flux density ratios and adopting the total flux density of the most recent measurement, to avoid problems due to source variability and inconsistencies in absolute flux density scale. The only discrepancy between C and A is at 1.7 GHz, where C is fainter than expected for a lensed image. One plausible explanation is free–free absorption due to ionized material $\lesssim 200 \text{ pc}$ from the lens galaxy nucleus, the approximate position of C. The dashed line is a two-parameter fit to the function $\mu_{\text{CA}} e^{-\tau_{\nu}}$, using a standard approximation²⁵ for free–free opacity, $\tau_{\nu} = 0.08235 T^{-1.35} \nu^{-2.1} E$, where ν is the frequency (in GHz) and T and E are the temperature (in Kelvin) and emission measure (the integral of n_e^2 over the path length s , in $\text{cm}^{-6} \text{ pc}$) of the ionized medium. The required opacity $\tau_{\nu} = 0.7$ at the lens-frame frequency $\nu \approx 3.4 \text{ GHz}$ could be produced by circumnuclear material with, for example, $T = 6,000$ and $E = 10^7$ ($n_e = 10^3 \text{ cm}^{-3}$, $s = 10 \text{ pc}$), which seems physically reasonable. For comparison, the emission measure is larger than in most individual Galactic H II regions ($E \approx 10^5\text{--}10^6$; for example see ref. 25), comparable to some ionized clouds around Sgr A ($E \approx 10^6\text{--}10^7$; for example see ref. 26), and smaller than observed around the nuclei of some radio galaxies ($E \approx 10^8$, for example see refs 27, 28). The free–free absorption hypothesis and the nature of the absorbing medium could be tested further with future low-frequency observations.

density apply, and firm up the correspondence with nearby galaxies.

Measuring correlated variability between C and another image would provide irrefutable evidence that C is a central image, should doubt still remain. Because of the different light-travel times, quasar flux variations should be seen first in A, then about 180 days later in B, and then about 9 hours later in C. The ratio of the delays would also further constrain mass models, although the C–B delay measurement would be difficult because of the faintness of the image and the necessarily fine time-sampling.

Finally, an additional image²¹ is predicted to occur over the entire allowed range of parameters of our models with central black holes. In general, the fourth image has <10% the flux of C, although there are very rare arrangements in which C represents a pair of merging images. It is separated by <20 mas from C, indicating the need for very-long-baseline interferometry. At present the strongest upper limit on the flux of an additional component is <30% that of C, based on the 5 σ limit at 8 GHz (ref. 15). Detection of the fourth image would be interesting because the separation and magnification ratio between the central images would provide a measurement of M_{BH} , rather than an upper limit. □

Received 26 September; accepted 3 December 2003; doi:10.1038/nature02279.

- Dyer, C. C. & Roeder, R. C. Possible multiple imaging by spherical galaxies. *Astrophys. J.* **238**, L67–L70 (1980).
- Burke, W. L. Multiple gravitational imaging by distributed masses. *Astrophys. J.* **244**, L1 (1981).
- Narasimha, D., Subramanian, K. & Chitre, S. M. “Missing image” in gravitational lens systems? *Nature* **321**, 45–46 (1986).
- Wallington, S. & Narayan, R. The influence of core radius on gravitational lensing by elliptical lenses. *Astrophys. J.* **403**, 517–529 (1993).
- Rusin, D. & Ma, C.-P. Constraints on the inner mass profiles of lensing galaxies from missing odd images. *Astrophys. J.* **549**, L33–L37 (2001).
- Evans, N. W. & Hunter, C. Lensing properties of cored galaxy models. *Astrophys. J.* **575**, 68–86 (2002).
- Keeton, C. R. Lensing and the centers of distant early-type galaxies. *Astrophys. J.* **582**, 17–29 (2003).
- Muñoz, J. A., Kochanek, C. S. & Keeton, C. R. Cusped mass models of gravitational lenses. *Astrophys. J.* **558**, 657–665 (2001).
- Lewis, G. F., Carilli, C., Papadopoulos, P. & Ivson, R. J. Resolved nuclear CO(1–0) emission in APM 08279+5255: gravitational lensing by a naked cusp? *Mon. Not. R. Astron. Soc.* **330**, L15–L18 (2002).
- Lewis, G. F. *et al.* Spatially resolved STIS spectra of the gravitationally lensed broad absorption line quasar APM082791+5255: the nature of component C and evidence for microlensing. *Mon. Not. R. Astron. Soc.* **334**, L7–L10 (2002).
- Gorenstein, M. V. *et al.* Detection of a compact radio source near the center of a gravitational lens: quasar image or galactic core? *Science* **219**, 54–56 (1983).
- Chen, G. & Hewitt, J. N. Multifrequency radio images of the Einstein ring gravitational lens MG 1131+0456. *Astrophys. J.* **106**, 1719–1728 (1993).
- Fassnacht, C. *et al.* B2045+265: A new four-image gravitational lens from CLASS. *Astron. J.* **117**, 658–670 (1999).
- Winn, J. N. *et al.* PMN J1632–0033: a new gravitationally lensed quasar. *Astron. J.* **123**, 10–19 (2002).
- Winn, J. N., Rusin, D. & Kochanek, C. S. Investigation of the possible third image and mass models of the gravitational lens PMN J1632–0033. *Astrophys. J.* **587**, 80–89 (2003).
- Winn, J. N., Kochanek, C. S., Keeton, C. R. & Lovell, J. E. J. The quintuple quasar: radio and optical observations. *Astrophys. J.* **590**, 26–38 (2003).
- Keeton, C. R. & Winn, J. N. The quintuple quasar: mass modeling and interpretation. *Astrophys. J.* **590**, 39–51 (2003).
- Narayan, R., Blandford, R. & Nityananda, R. Multiple imaging of quasars by galaxies and clusters. *Nature* **310**, 112–115 (1984).
- Subramanian, K., Chitre, S. M. & Narasimha, D. Microlensing of multiply imaged quasars: flux variations and vanishing of images. *Astrophys. J.* **289**, 37–51 (1985).
- Faber, S. *et al.* The centers of early-type galaxies with HST. IV. Central parameter relations. *Astron. J.* **114**, 1771–1796 (1997).
- Mao, S., Witt, H. J. & Koopmans, L. V. E. The influence of central black holes on gravitational lenses. *Mon. Not. R. Astron. Soc.* **323**, 301–307 (2001).
- Ferrarese, L. & Merritt, D. A fundamental relation between supermassive black holes and their host galaxies. *Astrophys. J.* **539**, L9–L12 (2000).
- Gebhardt, K. *et al.* A relationship between nuclear black hole mass and galaxy velocity dispersion. *Astrophys. J.* **539**, L13–L16 (2000).
- Tremaine, S. *et al.* The slope of the black hole mass versus velocity dispersion correlation. *Astrophys. J.* **574**, 740–753 (2002).
- Mezger, P. G. & Henderson, A. P. Galactic H II regions I. Observations of their continuum radiation at the frequency 5 GHz. *Astrophys. J.* **147**, 471–489 (1967).
- Becker, T. *et al.* Anatomy of the Sagittarius A complex. V. Interpolation of the Sgr A* spectrum. *Astron. Astrophys.* **307**, 450–458 (1996).
- Taylor, G. B. The symmetric parsec-scale jets of the radio galaxy Hydra A. *Astrophys. J.* **470**, 394–402 (1996).
- Walker, R. C. *et al.* VLBA absorption imaging of ionized gas associated with the accretion disk in NGC 1275. *Astrophys. J.* **530**, 233–244 (2000).
- Sheperd, M. C. in *Astronomical Data Analysis Software and Systems VI* (eds Hunt, G. & Payne, H.E.) 77–84 (ASP Conf. Ser. 125, ASP, San Francisco, 1997).

Acknowledgements We thank S. Doeleman, D. Harris and P. Schechter for discussions, and J. Bullock for comments on the manuscript. J.N.W. acknowledges the support of the National Science Foundation (NSF) through an Astronomy and Astrophysics Postdoctoral Fellowship. The VLA is part of the National Radio Astronomy Observatory, an NSF facility operated under cooperative agreement by Associated Universities, Inc.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.N.W. (jwinn@cfa.harvard.edu).

A high-speed silicon optical modulator based on a metal–oxide–semiconductor capacitor

Ansheng Liu¹, Richard Jones¹, Ling Liao¹, Dean Samara-Rubio¹, Doron Rubin², Oded Cohen², Remus Nicolaescu¹ & Mario Paniccia¹

¹Intel Corporation, 2200 Mission College Blvd, CHP3-109, Santa Clara, California 95054, USA

²Intel Corporation, S. B. I. Park Har Hotzvim, Jerusalem, 91031, Israel

Silicon has long been the optimal material for electronics, but it is only relatively recently that it has been considered as a material option for photonics¹. One of the key limitations for using silicon as a photonic material has been the relatively low speed of silicon optical modulators compared to those fabricated from III–V semiconductor compounds^{2–6} and/or electro-optic materials such as lithium niobate^{7–9}. To date, the fastest silicon-waveguide-based optical modulator that has been demonstrated experimentally has a modulation frequency of only ~20 MHz (refs 10, 11), although it has been predicted theoretically that a ~1-GHz modulation frequency might be achievable in some device structures^{12,13}. Here we describe an approach based on a metal–oxide–semiconductor (MOS) capacitor structure embedded in a silicon waveguide that can produce high-speed optical phase modulation: we demonstrate an all-silicon optical modulator with a modulation bandwidth exceeding 1 GHz. As this technology is compatible with conventional complementary MOS (CMOS) processing, monolithic integration of the silicon modulator with advanced electronics on a single silicon substrate becomes possible.

Modulation of the refractive index in silicon can be achieved via the free carrier plasma dispersion effect^{14,15}. In contrast to the conventional way of inducing a free carrier density change in silicon (through injection or depletion of electrons and holes in the intrinsic region of a forward biased silicon p–i–n diode^{10–13} or a three-terminal device^{16,17}), we demonstrate the use of an MOS capacitor phase shifter to realize the charge density modulation. Under accumulation conditions, the majority carriers in the silicon waveguide modify the refractive index so that phase shift is induced in the optical mode. The advantage of using a MOS capacitor phase shifter is the high achievable modulation speed, as there are no slow carrier generation and/or recombination processes involved in the accumulation operation.

Figure 1 shows a diagram of the cross-sectional view of our silicon-waveguide-based MOS capacitor phase shifter. It comprises an n-type doped crystalline silicon slab (the silicon layer of the ‘silicon-on-insulator’ wafer) and a p-type doped polysilicon rib with a gate oxide sandwiched between them. The polysilicon is formed by first depositing amorphous silicon using low pressure chemical vapour deposition at 550 °C. This amorphous layer is then

crystallized into polysilicon using a high temperature anneal. The annealing process is designed to minimize polysilicon grain boundaries, which can cause optical loss and limit the electric activation of dopants^{18,19}.

In order to minimize the optical absorption by metal contacts, we used a wide ($\sim 10.5 \mu\text{m}$) polysilicon layer on the top of the oxide layers on both sides of the polysilicon rib. Aluminium contacts are deposited on top of this polysilicon layer. By placing the metal contacts to the side rather than on top of the rib waveguide (as often done in silicon photonic devices¹⁰), the optical absorption due to the metal contacts is almost eliminated. The oxide regions on either side of the rib maintain optical confinement and prevent optical field penetrating into the areas where the metal contacts are located. To make ohmic contacts to the metal, both the crystalline silicon slab and polysilicon have a surface doping concentration of $1 \times 10^{19} \text{ cm}^{-3}$. Our MOS phase modulators were fabricated in an existing Intel CMOS production facility, demonstrating compatibility with standard CMOS processing.

In the rib waveguide, optical confinement in the horizontal direction is achieved by a conventional rib structure, while vertical optical confinement is due to the $\sim 0.375\text{-}\mu\text{m}$ buried oxide and an oxide cover (not shown in Fig. 1). Modelling and testing confirm that the waveguide phase shifter shown in Fig. 1, containing a thin gate oxide and a thin polysilicon top layer, is a single-mode device at wavelengths around $1.55 \mu\text{m}$. Simulation also shows that $\sim 9\%$ of optical power for the TE mode is confined in the polysilicon region. The presence of a horizontal gate oxide in the waveguide induces strong polarization dependence of the modal characteristics. Phase modulation efficiency for TE polarized light is much larger (a factor of 7) than that for TM polarized light. This implies that the current phase shifter is inherently a polarization sensitive device, in common with commercially available LiNbO_3 modulators⁷⁻⁹. In all the measurements presented here, TE polarized light was used.

In the accumulation operation, the n-type silicon in the phase shifter is grounded and a positive drive voltage, V_D , is applied to the p-type polysilicon. When a positive voltage V_D is applied to the device, a thin charge layer is accumulated on both sides of the gate oxide. The voltage-induced charge density change ΔN_e (for electrons) and ΔN_h (for holes) is related to the drive voltage by²⁰:

$$\Delta N_e = \Delta N_h = \frac{\epsilon_0 \epsilon_r}{et_{\text{ox}} t} [V_D - V_{\text{FB}}] \quad (1)$$

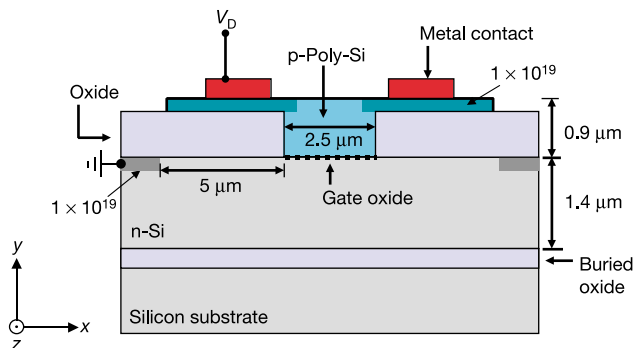


Figure 1 Schematic diagram showing the cross-sectional view of a MOS capacitor waveguide phase shifter using 'silicon-on-insulator' technology. The n-type doped crystalline silicon layer thickness is $\sim 1.4 \mu\text{m}$, and the p-type doped polysilicon thickness at the centre of the waveguide is $\sim 0.9 \mu\text{m}$. The gate oxide thickness is $120 \text{ \AA}$. The polysilicon rib and the gate oxide widths are both $\sim 2.5 \mu\text{m}$. The n-type silicon has an active doping concentration of $\sim 1.7 \times 10^{16} \text{ cm}^{-3}$ and the p-type polysilicon has an active doping concentration of $\sim 3 \times 10^{16} \text{ cm}^{-3}$. These doping concentrations were chosen to produce a modulation speed of the phase shifter of above 1 GHz, as the speed of the device depends on silicon and polysilicon resistances. A surface doping density of $1 \times 10^{19} \text{ cm}^{-3}$ was designed to minimize the metal-semiconductor contact resistance.

where ϵ_0 and ϵ_r are the vacuum permittivity and low-frequency relative permittivity of the oxide, e is the electron charge, t_{ox} is the gate oxide thickness, t is the effective charge layer thickness, and V_{FB} is the flat band voltage of the MOS capacitor. In the presence of free carriers in silicon, both refractive index and optical absorption are changed^{14,15}. At a wavelength of $1.55 \mu\text{m}$, the refractive index changes caused by accumulated electrons and holes, which were obtained from experimental absorption spectra through Kramers-Kronig analysis²¹, are given by:

$$\Delta n_e = -8.8 \times 10^{-22} \Delta N_e \quad (2)$$

$$\Delta n_h = -8.5 \times 10^{-18} (\Delta N_h)^{0.8} \quad (3)$$

where electron and hole density changes are in units of cm^{-3} . The change in refractive index results in a phase shift $\Delta\phi$ in the optical mode given by:

$$\Delta\phi = \frac{2\pi}{\lambda} \Delta n_{\text{eff}} L \quad (4)$$

where L is the active length of the phase shifter, λ is the wavelength of light in free space, and Δn_{eff} is the effective index change in the waveguide, which is the difference between the effective indices of the waveguide phase shifter before and after charge accumulation.

To evaluate the phase shift, we first determine the refractive index change due to the voltage induced free carrier density through equations (1)–(3). Using a two-dimensional device simulation package DESSIS, we modelled the induced charge density as a function of the drive voltage. This device simulation tool uses the 'box discretization' method²² to solve the coupled Poisson equation and electron and hole continuity equations that govern charge transport in semiconductor devices. This is used to predict the electrical characteristics associated with specified physical structures and conditions. In addition, the simulation tool incorporates advanced semiconductor physics models to account for carrier statistics, doping and field dependent carrier mobility, and doping dependent carrier recombination lifetimes. The simulation results suggest that the flat band voltage of our devices is $V_{\text{FB}} = 1.25 \text{ V}$, and the effective charge layer thickness is $t = 10 \text{ nm}$, as defined in equation (1). Subsequently, the refractive index change in the accumulated charge layers is used to calculate the effective index change of the waveguide mode. As the charge layer thickness is small

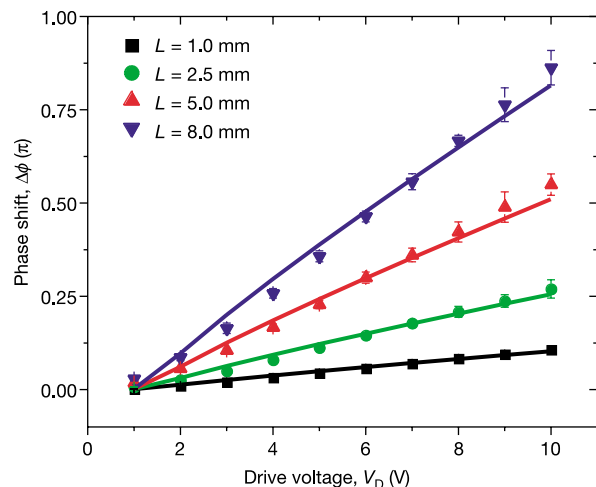


Figure 2 Phase shift $\Delta\phi$ versus drive voltage V_D of the MOS capacitor phase shifter in Fig. 1 at a wavelength of $\lambda = 1.55 \mu\text{m}$ for different phase shifter lengths. The symbols represent the measured phase shifts, and the solid lines are the simulated phase shifts.

(10 nm) and there is a thin gate oxide (12 nm) in the waveguide, care must be taken to simulate accurately the effective index change of the guided mode. In our simulations, we used a fully vectorial Maxwell waveguide solver based on the film mode matching method²³ to simulate the waveguide mode. After calculating the effective index change Δn_{eff} , the phase shift is then obtained from equation (4).

Before showing experimental results of integrated Mach–Zehnder interferometer (MZI) modulators, we characterize the phase shift for an individual phase shifter. Figure 2 shows the measured and modelled phase shift as a function of the drive voltage for different phase shifter lengths. The phase measurement was performed in a free space interferometer. As can be seen from Fig. 2, there is a good agreement between simulation and measurement for different phase shifter lengths, that is, $L = 1, 2.5, 5$ and 8 mm. The measured phase shift is almost linearly dependent upon the drive voltage. The data also confirm the linear dependence of the phase shift on the phase shifter length for a given drive voltage, as expected from equation (4). As a figure of merit, the product $V_{\pi}L$ can be determined from the measured phase shift, where V_{π} is the drive voltage swing (or $V_D - V_{\text{FB}}$) required for π phase shift and L is the device length. Using the data in Fig. 2, and taking into account the flat band voltage of $V_{\text{FB}} = 1.25$ V, we obtain a $V_{\pi}L$ product for our current device of ~ 8 V cm.

To convert the phase modulation into an intensity modulation, we fabricated an asymmetric MZI by inserting two identical MOS capacitor phase shifters in the two arms of the MZI, as shown schematically in Fig. 3a. The asymmetric MZI has an optical path length difference of ~ 16.7 μm between the two arms. A Y junction is used to split and combine the optical beam in the MZI. To characterize the performance of our silicon MZI modulator, we measured the optical output intensity as a function of drive voltage for a device having a phase shifter length of 10 mm. In the measurement, we used an input light wavelength of 1.54 μm and applied the drive voltage to one of the two phase shifters in the MZI. A typical result is given in Fig. 3b, which shows that the extinction ratio of the MZI is more than 16 dB with an applied peak-to-trough

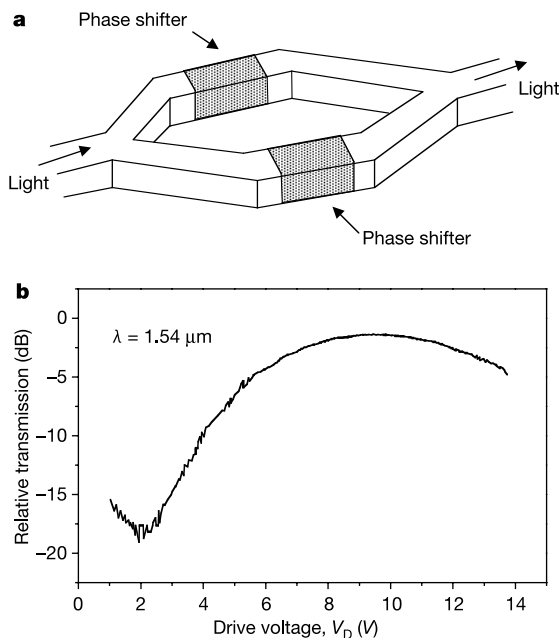


Figure 3 Integrated silicon optical modulator and measured drive-voltage-dependent output optical intensity. **a**, Schematic diagram showing an asymmetric Mach–Zehnder interferometer (MZI) containing two identical MOS capacitor phase shifters in the two arms. **b**, Measured relative transmission of a MZI as a function of applied voltage at a wavelength of 1.54 μm .

voltage of ~ 7.7 V. Our MZI modulator has an insertion loss of ~ 15.3 dB when the MZI is in the ‘on’ state, in which the optical output intensity of the MZI reaches its maximum value. This insertion loss includes a ~ 4.3 dB per interface coupling loss to and from the waveguide, using lensed fibres, and a ~ 6.7 dB on-chip loss that is mainly attributed to the doped polysilicon phase shifter and undoped polysilicon waveguide. The waveguide loss was measured by use of the cutback method¹⁸.

We characterized the high-frequency performance of our silicon modulator with both a 0.18-V r.m.s. sinusoidal source and a 3.5-V digital pulse pattern. In both of the measurements, we used an asymmetric MZI containing a single 2.5-mm phase shifter. For the 0.18-V signal measurement, we used an input light wavelength of 1.558 μm and used lensed fibres to couple into and out of the device under test. To maximize the optical a.c. amplitude (the modulation depth) of the MZI, we applied a d.c. bias of 3 V to the phase shifter to bias the MZI to a high-sensitivity region. We then applied the a.c. drive voltage to the phase shifter. The optical output intensity of the MZI was detected using a high-frequency photoreceiver with a bandwidth of 15 GHz. To confirm that the high-frequency electrical a.c. signal was appropriately applied to the phase shifter, we also monitored the on-chip voltage as a function of the frequency, while measuring the optical intensity modulation. The measured frequency dependence of the photoreceiver r.m.s. voltage output signal and the on-chip r.m.s. voltage is shown in Fig. 4a. Because our device is capacitive, the on-chip voltage experiences a high-frequency roll-off, which is governed by the output impedance of the drive circuitry and the capacitance of the phase shifter. To obtain the frequency dependence of the optical response of the MZI modulator at a constant drive voltage, we normalized the photoreceiver output voltage by the on-chip drive voltage at the applied frequency. The resulting response is shown in Fig. 4b. We see from Fig. 4b an unambiguous demonstration of the fact that the MZI modulator has a 3-dB roll-off frequency in excess of the benchmark figure of 1 GHz. Such a modulation frequency is two orders of magnitude

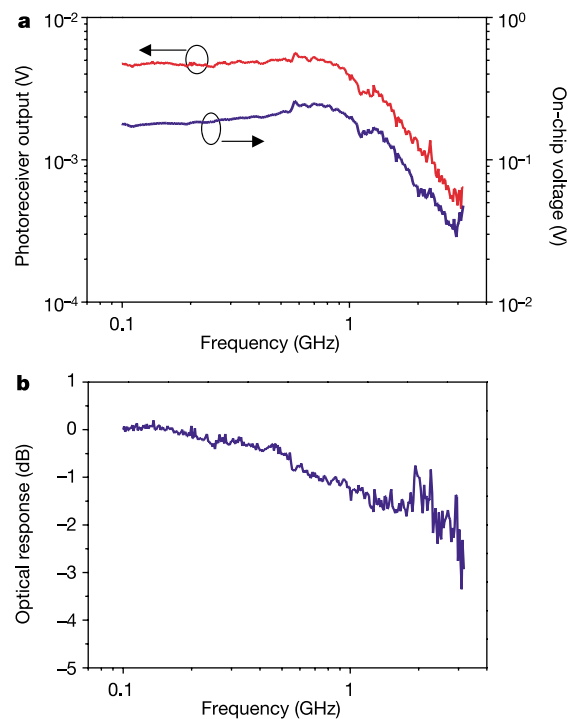


Figure 4 Frequency dependence of the optical response of a silicon MZI modulator. **a**, Measured photoreceiver output and on-chip r.m.s. voltage as a function of frequency for a MZI modulator containing a single 2.5-mm-long MOS capacitor phase shifter in one arm. **b**, Optical response of the MZI modulator as a function of frequency.

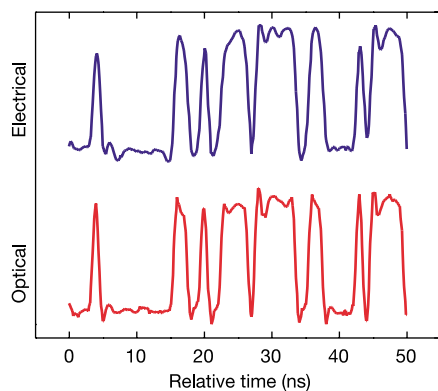


Figure 5 Pseudorandom bit sequence of a silicon Mach-Zehnder interferometer modulator containing a single 2.5-mm-long MOS capacitor phase shifter in one arm at a data bit rate of 1 Gbit s^{-1} .

higher than that of traditional current-injection-based p-i-n diode devices^{10,11}.

We also investigated the data transmission performance of the MZI modulator with a digital pulse drive voltage. We applied 3.5-V digital pulse pattern with a d.c. bias of 3 V to the phase shifter. With a pseudorandom electrical data input, we detected the corresponding MZI optical output by using a high-frequency photoreceiver. In Fig. 5 we show a 1 Gbit s^{-1} pseudorandom electrical bit sequence input and the corresponding optical output of our modulator. The data show that the optical signal faithfully reproduces the 1 Gbit s^{-1} electrical data stream.

As mentioned above, our current MZI modulator has an on-chip loss of $\sim 6.7 \text{ dB}$, which is primarily due to the doped and undoped polysilicon regions in the waveguide. Because the polysilicon has a much larger optical loss than single-crystal silicon^{18,19}, we could significantly reduce the present modulator loss by replacing the polysilicon region with single-crystal silicon. For the device in Fig. 1, modelling suggests that we could reduce the on-chip loss by $\sim 5 \text{ dB}$ by this method. Using the epitaxial lateral overgrowth technique²⁴, single-crystal silicon can be fabricated over the gate oxide. Investigation of the replacement of polysilicon by single-crystal silicon is under way. Furthermore, because the phase shifter loss depends on the device length, that is, L_π (the active waveguide length required for π phase shift), increasing the phase modulation efficiency (or decreasing the $V_\pi L$ product) by reducing the waveguide dimension and making the gate oxide thinner not only reduces the device size but also the optical loss. Scaling down the waveguide helps the high-frequency operation, as the capacitance is reduced by the device size shrinkage. In addition, reducing the gate oxide thickness and commensurately reducing the drive voltage also reduces the high-frequency power dissipation of the modulator, as with the conventional CMOS circuits²⁵. We could also design and fabricate a graded doping profile in the y direction (Fig. 1) in the waveguide phase shifter to further reduce the optical loss, while still maintaining the device speed. For example, we could design a phase shifter with higher doping densities in the areas close to the gate oxide and metal contacts, but with lower doping concentrations in the rest of the waveguide. □

Received 22 September; accepted 18 December 2003; doi:10.1038/nature02310.

1. Soref, R. A. Silicon-based optoelectronics. *Proc. IEEE* **81**, 1687–1706 (1993).
2. Zucker, J. E., Jones, K. L., Miller, B. I. & Koren, U. Miniature Mach-Zehnder InGaAsP quantum well waveguide interferometers for $1.3 \mu\text{m}$. *IEEE Photon. Technol. Lett.* **2**, 32–34 (1990).
3. Cites, J. S. & Ashley, P. R. High-performance Mach-Zehnder modulators in multiple quantum well GaAs/AlGaAs. *J. Lightwave Technol.* **12**, 1167–1173 (1992).
4. Fetterman, M., Chao, C.-P. & Forrest, S. R. Fabrication and analysis of high-contrast InGaAsP-InP Mach-Zehnder modulators for use at $1.55\text{-}\mu\text{m}$ wavelength. *IEEE Photon. Technol. Lett.* **8**, 69–71 (1996).
5. Leclerc, O. et al. Polarisation-independent InP push-pull Mach-Zehnder modulator for 20 Gbit/s soliton regeneration. *Electron. Lett.* **34**, 1011–1013 (1998).

6. Ido, T. et al. Ultra-high-speed multiple-quantum-well electro-absorption optical modulators with integrated waveguides. *J. Lightwave Technol.* **14**, 2026–2034 (1996).
7. Fujiwara, T., Watanabe, A. & Mori, H. Measurement of uniformity of driving voltage in Ti:LiNbO₃ waveguides using Mach-Zehnder interferometers. *IEEE Photon. Technol. Lett.* **2**, 260–261 (1990).
8. Wooten, E. L. et al. A review of lithium niobate modulators for fiber-optic communications systems. *IEEE J. Select. Topics Quant. Electron.* **6**, 69–82 (2000).
9. Howerton, M. M., Moeller, R. P., Greenblatt, A. S. & Krahenbuhl, R. Fully packaged, broad-band LiNbO₃ modulator with low drive voltage. *IEEE Photon. Technol. Lett.* **12**, 792–794 (2000).
10. Tang, C. K. & Reed, G. T. Highly efficient optical phase modulator in SOI waveguides. *Electron. Lett.* **31**, 451–452 (1995).
11. Dainesi, P. et al. CMOS compatible fully integrated Mach-Zehnder interferometer in SOI technology. *IEEE Photon. Technol. Lett.* **12**, 660–662 (2000).
12. Png, C. E., Reed, G. T., Atta, R. M. H., Ensell, G. J. & Evans, A. G. R. Development of small silicon modulators in silicon-on-insulator (SOI). *Proc. SPIE* **4997**, 190–197 (2003).
13. Irace, A., Breglio, G. & Cutolo, A. All-silicon optoelectronic modulator with 1 GHz switching capability. *Electron. Lett.* **39**, 232–233 (2003).
14. Soref, R. A. & Lorenzo, P. J. All-silicon active and passive guided-wave components for $\lambda = 1.3$ and $1.6 \mu\text{m}$. *IEEE J. Quant. Electron.* **QE-22**, 873–879 (1986).
15. Soref, R. A. & Bennett, B. R. Electrooptical effects in silicon. *IEEE J. Quant. Electron.* **QE-23**, 123–129 (1987).
16. Cutolo, A., Iodice, M., Spirito, P. & Zeni, L. Silicon electro-optic modulation based on a three terminal device integrated in a low-loss single-mode SOI waveguide. *J. Lightwave Technol.* **15**, 505–518 (1997).
17. Sciuto, A., Libertino, S., Alessandria, A., Coffa, S. & Coppola, G. Design, fabrication, and testing of an integrated Si-based light modulator. *J. Lightwave Technol.* **21**, 228–235 (2003).
18. Liao, L. et al. Optical transmission losses in polycrystalline silicon strip waveguides: effects of waveguide dimensions, thermal treatment, hydrogen passivation, and wavelength. *J. Electron. Mater.* **29**, 1380–1386 (2000).
19. Liao, L. *Low Loss Polysilicon Waveguides for Silicon Photonics*. Thesis, MIT (1997).
20. Sze, S. M. *Physics of Semiconductor Devices* 2nd edn (Wiley, New York, 1981).
21. Soref, R. A. & Bennett, B. R. Kramers-Kronig analysis of electro-optical switching in silicon. *Proc. SPIE* **704**, 32–37 (1986).
22. Bank, R. E., Rose, D. J. & Fichtner, W. Numerical methods for semiconductor device simulation. *IEEE Trans. Electron. Dev.* **ED-30**, 1031–1041 (1983).
23. Sudbo, A. S. Numerically stable formulation of the transverse resonance method for vector mode-field calculations in dielectric waveguides. *IEEE Photon. Technol. Lett.* **5**, 342–344 (1993).
24. Ahmed, S. S., Denton, J. P. & Neudeck, G. W. Nitrided thermal SiO₂ for use as top and bottom gate insulators in self-aligned double gate silicon-on-insulator metal-oxide-semiconductor field effect transistor. *J. Vac. Sci. Technol. B* **19**, 800–806 (2001).
25. Davari, B., Dennard, R. H. & Shahidi, G. G. CMOS scaling for high performance and low power—The next ten years. *Proc. IEEE* **83**, 595–606 (1995).

Acknowledgements We thank M. Salib for help in process design and optical loss testing; M. Morse, A. Barkai, S. Tubul and D. Tran for technical assistance in device fabrication; A. Alduino for backend processing; and S. Koehl for data collection software. Special thanks go to D. Elqaq, M. Gill, S. Pang and B. Venkateshwaran for contributions during the early stages of this research. Finally, we thank G. Reed for discussions.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.L. (ansheng.liu@intel.com).

A 1.7-kilobase single-stranded DNA that folds into a nanoscale octahedron

William M. Shih¹, Joel D. Quispe² & Gerald F. Joyce¹

¹Departments of Chemistry and Molecular Biology and The Skaggs Institute for Chemical Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA

²Department of Cell Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA

Molecular self-assembly offers a means of spontaneously forming complex and well-defined structures from simple components. The specific bonding between DNA base pairs has been used in this way to create DNA-based nanostructures and to direct the assembly of material on the subnanometre to micrometre scale^{1,2}. In principle, large-scale clonal production of suitable DNA sequences and the directed evolution of sequence lineages

towards optimized behaviour³ can be realized through exponential DNA amplification by polymerases. But known examples of three-dimensional geometric DNA objects^{4–6} are not amenable to cloning because they contain topologies that prevent copying by polymerases^{1,2,7}. Here we report the design and synthesis of a 1,669-nucleotide, single-stranded DNA molecule that is readily amplified by polymerases and that, in the presence of five 40-mer synthetic oligodeoxynucleotides, folds into an octahedron structure by a simple denaturation–renaturation procedure. We use cryo-electron microscopy to show that the DNA strands fold successfully, with 12 struts or edges joined at six four-way junctions to form hollow octahedra approximately 22 nanometres in diameter. Because the base-pair sequence of individual struts is not repeated in a given octahedron^{8,9}, each strut is uniquely addressable by the appropriate sequence-specific DNA binder.

Rigid DNA structures can be constructed from double-helical struts that are linked at flexible joints (for example, branched junctions) by arrangement of the struts in a continuously triangulated frame². DNA tetrahedra⁶ have an architecture that resists deformation, whereas non-triangulated objects, such as DNA cubes⁴ and truncated octahedra⁵, do not. A DNA regular octahedron, also continuously triangulated, has been proposed previously¹⁰. Some viruses fold parts of their single-stranded RNA genomes into triangulated (for example, icosahedral¹¹) or untriangulated (for example, dodecahedral¹²) geometric shapes, with encapsulating protein shells enforcing these RNA structures.

The DNA octahedron reported here consists of five double-crossover¹⁰ (DX) struts and seven paranemic-crossover¹³ (PX) struts, joined at six four-way junctions (Fig. 1a). DX and PX motifs

have both been modelled as pairs of double helices that are arranged in a side-by-side manner. Each of the twelve struts of the octahedron contributes one double helix to a 'core' layer and the other to a concentric 'peripheral' layer. The four-way junctions connect only the core-layer double helices. Each four-way junction displays on its concave face the minor grooves of its four proximally surrounding base pairs. All four strands at all six junctions contain two unpaired thymidine residues at the crossover point to allow some flexibility for assembly.

The core-layer double helix of each of the twelve struts contains 40 base pairs, corresponding to roughly four turns of DNA and a length of ~14 nm. For eleven of the struts, the peripheral-layer double helix contains 30 or 32 base pairs and is capped at both helical ends by a hairpin loop of four thymidine residues. The twelfth strut is slightly longer, containing 35 base pairs, and is capped at only one end. DX motifs have been shown to be about twice as stiff as standard duplex DNA¹⁴. Assuming that PX motifs have similar rigidity, each of the struts would be expected to have a stiffness corresponding to a rod that is roughly one-eighth of a persistence length. Thus the folded octahedron, in which all twelve struts come together, is expected to be a highly rigid object.

Each DX motif contains 20 base pairs between the two crossover junctions, where the strands exchange between the two component double helices. The crossover junctions are flanked on the core-layer helix by eight base pairs on one side and twelve base pairs on the other side, and are flanked on the peripheral-layer helix by either five or six base pairs on each side. Each PX motif contains six crossover junctions separated by three major-groove spacings of six base pairs alternating with two minor-groove spacings of four base pairs. The outermost crossover junctions are flanked on the core-layer helix by seven base pairs on either side, and are flanked on the peripheral-layer helix by three base pairs on either side.

The DNA octahedron was folded from a mixture of a 1,669-nucleotide 'heavy chain' and five 40-nucleotide 'light chains' by heat denaturation followed by successive cooling steps. Folding was designed to occur in two stages. First, the heavy chain and the five light chains associate stoichiometrically and collapse into a branched-tree structure (Fig. 1b). Binding of the heavy and light chains forms double crossovers that provide five of the twelve struts of the target structure. This intermediate state has fourteen terminal branches, each corresponding to a half-strut. The terminal branches are unique 76-nucleotide loops, each with sequence complementarity (in the PX sense) to one and only one other loop sequence (Fig. 1c). In the second stage of folding, conjugate terminal branches associate to form the remaining seven struts. The order of formation of the struts should not make a difference in achieving the final structure.

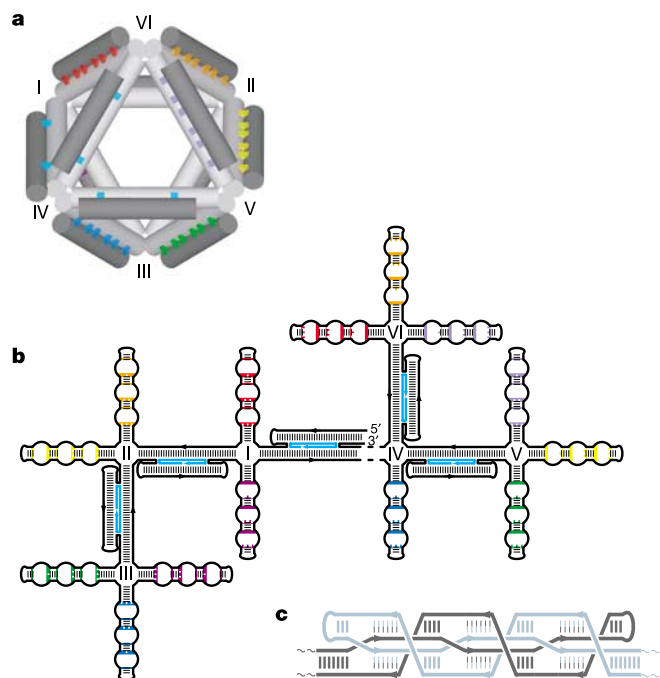


Figure 1 Design of the DNA octahedron. **a**, Three-dimensional structure involving twelve struts (octahedron edges) connected by six flexible joints (octahedron vertices). Five of the struts are DX motifs (cyan) and seven are PX motifs (rainbow colours). The joints are four-way junctions that connect the core-layer double helices of each strut. **b**, Secondary structure of the branched-tree folding intermediate. The structure consists of a single heavy chain (black) and five unique light chains (cyan). Like colours indicate half-PX loops whose sequence-specific cross-association generates a strut that serves as an edge of the DNA octahedron. Coloured stripes coincide with strand crossover positions. Folding to the structure in the upper left is complete when all seven PX struts have formed. **c**, Schematic of a PX strut.

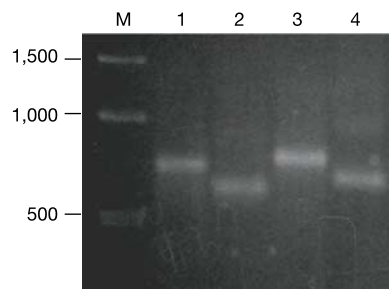


Figure 2 Gel-shift analysis of folding of the DNA octahedron. Lane M, marker lane with DNA size standards (number of base pairs indicated at the left). Lane 1, heavy chain folded in the absence of Mg^{2+} . Lane 2, heavy chain folded in the presence of Mg^{2+} . Lane 3, heavy and light chains folded in the absence of Mg^{2+} . Lane 4, heavy and light chains folded in the presence of Mg^{2+} . Samples were electrophoresed at 100 V at 4 °C in a 2% agarose gel containing 2 mM $MgCl_2$, 0.5 $\mu g\ ml^{-1}$ ethidium bromide, 45 mM Tris base, and 45 mM boric acid (pH 8.0).

The DNA octahedron reported here contains no catenations or knots. Therefore no covalent bonds need to be formed or broken to attain the final structure. It was suggested previously that catenated DNA objects might be formed from a single DNA strand by topoisomerization to establish the required knots followed by restriction digestion and ligation to achieve conversion to a catenane⁷, although this has not been demonstrated experimentally. An uncatenated structure, as exemplified by the present study, may be less robust because no covalent bonds need to be broken to sever the edges. Yet this structure was preferred because it can be formed by a simple folding protocol.

An initial test for proper folding was provided by a gel mobility-shift assay (Fig. 2). During folding in the absence of Mg^{2+} , the half-PX terminal branches should be unable to associate, leaving the DNA object with the extended branched-tree structure. During folding in the presence of Mg^{2+} , the terminal branches should associate to form PX struts, resulting in a more compact structure. Consistent with these expectations, the DNA exhibited an increased gel mobility when folded in the presence of Mg^{2+} compared to its absence. If any of the fourteen half-PX loops had not been satisfied by an intramolecular partner, then they would have been available for pairing with a separate molecule. About half of the material was observed to be monomeric after folding in the presence of Mg^{2+} (Fig. 2, lane 4), suggesting that most of the half-PX loops were satisfied intramolecularly or were otherwise unavailable for intermolecular pairing.

The overall structure of the DNA octahedron was revealed using cryo-electron microscopy and single-particle reconstruction techniques^{15–17}. Figure 3a shows a representative cryo-electron micrograph, demonstrating numerous octahedral-shaped objects of the

expected size. Some particles in the micrographs do not appear to be properly folded, although the low signal-to-noise ratio of the images makes it difficult to determine what fraction of the particles have adopted the target structure. Distorted particles in the image field could be the result either of misfolding or of rearrangement during preparation for cryo-freezing, perhaps induced by adhesion to the supporting carbon substrate. The best estimate for the fraction of properly folded particles comes from the gel mobility-shift data (Fig. 2), which suggests that about half of the material is properly folded.

Figure 3b shows a three-dimensional map of the structure of the DNA octahedron, reconstructed from 961 particles. The signal-to-noise ratio was too low to allow meaningful reconstruction without imposing some form of symmetry. Initially, four-fold rotational symmetry was imposed, which produced a map with an octahedral shape. Then a second reconstruction was performed imposing octahedral symmetry, resulting in a better-defined structure. Decoy reconstructions were performed imposing other symmetries (for example, D3, C5, D5, C6), but all of these produced a map whose projections match poorly with the corresponding class averages, whereas there is a close match for the octahedral map (see Supplementary Information). The folded object has the overall shape of an octahedron, but at a fine level it is not perfectly regular because each strut has a different DNA sequence.

Figure 3c shows raw images of representative particles, each accompanied by a corresponding projection of the computed map. The density along the struts in this map demonstrates a paddle shape that is consistent in dimensions with two adjacent double helices having an overall cross-section of about $2\text{ nm} \times 4\text{ nm}$. The struts appear to be pinned at four-way joints connected on the inside corners of each strut, consistent with the designed structure of four-way junctions that join core-layer double helices and peripheral-layer double helices that lie outside the framework defined by the core layer. The cavity enclosed by the struts is large enough to accommodate a sphere having a diameter up to 14 nm, while the triangular opening on each face is large enough to allow passage of a sphere having a diameter of up to 8 nm.

Formation of an octahedron structure from the branched-tree intermediate could in principle follow an erroneous course, in which the four-way junctions fold with the major instead of minor grooves of the surrounding base pairs on the concave face of each junction. In this misfolded isomer, the double helices that are not connected by four-way junctions (corresponding to the peripheral-layer helices of the correct isomer) would lie on the inside of the octahedron. Severe steric clashes seem to prevent this alternative isomer from forming.

The assembly strategy presented here could be modified to produce catenated structures, such as the DNA cube⁴ reported previously, by substituting hairpin loops that encode restriction sites for the half-PX loops. Cleavage of the restriction sites would produce sticky ends that could be ligated to their appropriate partners. The requirement for light chains could be eliminated by replacing DX struts with simple duplex struts. DX struts were preferred here to produce an octahedron with greater rigidity and pseudosymmetry.

This study has demonstrated that a long single-stranded DNA, accompanied by five synthetic 40-mer oligodeoxynucleotides, can fold without knotting into a DNA cage having the form of an octahedron. The heavy-chain DNA can be amplified in the context of a bacterial plasmid and later excised, allowing its clonal production. The DNA octahedron can be modified in a modular fashion by extending the ends of any of the peripheral double helices or by making insertions into the struts. Sequence variants, perhaps displaying variable loops at multiple locations on the DNA scaffold, could be used to initiate directed evolution studies.

Manipulation of the basic octahedron structure could produce a family of scaffold variants whose function is to position other

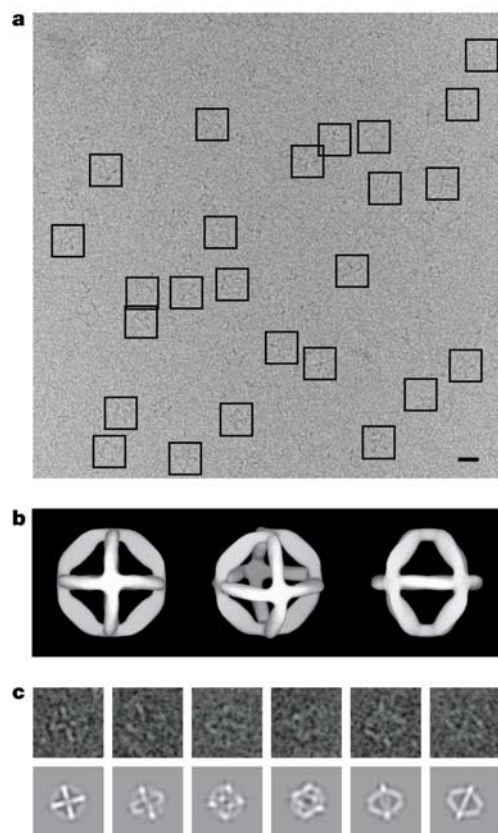


Figure 3 Visualization of the DNA octahedron structure by cryo-electron microscopy. **a**, Representative cryo-electron micrograph, with 25 individual DNA particles boxed. Scale bar, 20 nm. **b**, Three views of the three-dimensional map generated from single-particle reconstruction of the DNA octahedron. **c**, Raw images of individual particles and corresponding projections of the three-dimensional map.

material at uniquely defined locations. Just as messenger RNAs encode the one-dimensional positioning of transfer-RNA adaptors along their lengths, rigid DNA nanostructures could encode the positioning of sequence-specific DNA binders along their three-dimensional frameworks. DNA nanostructures, therefore, can be regarded as devices that use molecular recognition to translate sequence information into three-dimensional positions. The demonstration of a clonable DNA octahedron is a step towards making the translation of DNA sequence information into position more practical and more versatile. □

Methods

Design of sequences

PX sequences were chosen as will be described elsewhere (W.M.S. & G.F.J., manuscript in preparation). DX sequences were chosen arbitrarily, using the model junction J1¹⁸ for the crossover junctions. The DX sequences were examined to eliminate any seven-nucleotide repeats. Folding of the assembled heavy chain was analysed using MFOLD¹⁹. Iterative changes were made to the sequence until the target branched-tree structure was achieved as the lowest-energy output.

Folding conditions

Folding was achieved by incubation at 90 °C for 5 min, 65 °C for 20 min, 55 °C for 20 min, 45 °C for 20 min, and 37 °C for 30 min. The folding mixture contained 100 nM heavy chain, 400 nM each of the five light chains, 10 mM MgCl₂, 50 mM NaCl, and 40 mM EPPS (pH 7.5).

Synthesis, cloning, and purification of single-stranded DNA

The 1,669-nucleotide heavy chain was constructed by polymerase chain reaction (PCR)-based assembly²⁰, using synthetic oligodeoxynucleotides prepared by standard phosphoramidite chemistry on a PerSeptive Biosystems Expedite 8900 automated DNA synthesizer. The PCR products were gel purified and ligated into pBS II KS⁺ plasmids, which were transformed into *Escherichia coli* (see Supplementary Information for additional details). Multiple clones were selected and sequenced to identify error-free clones. An *EcoR* V site was placed upstream and an *N.Bbv* C IA (nicking endonuclease) site was placed downstream of the heavy-chain sequence using the Stratagene Quikchange cloning kit. The plasmid was amplified in *E. coli* and purified using the QIAGEN QiaQuick ion-exchange chromatography system. The desired single strand was obtained by excision with *EcoR* V and *N.Bbv* C IA, then purified by electrophoresis in a 1.5% alkaline agarose gel, and further purified by electrophoresis in a 1.5% native agarose gel. The light chains were prepared by chemical synthesis. The sequences of all oligodeoxynucleotides used in this study are provided in the Supplementary Information.

Electron microscopy

Cryo-electron microscopy was performed as described¹⁵. A 4-μl drop of folded DNA, concentrated to 500 nM by microdialysis, was placed on a 300-mesh copper grid covered with a continuous carbon substrate and glow-discharged in amyl amine. The grid was blotted for ~2 s with Whatman filter paper no. 1 and plunge-frozen into liquid ethane. The specimen was observed at -176 °C using a Gatan 626 cryo-stage on a Philips Tecnai F20 FEG operating at 120 kV. Paired images at approximately 0.6 and 2.0 μm defocus were recorded on a Tietz 2k × 2k charge-coupled device (CCD) camera under low-dose conditions using the Leginon²¹ automated data collection software system. The magnification was ×62,000, resulting in a pixel size of 2.24 Å. The best 961 images of boxed particles at 2.0-μm defocus were used for three-dimensional reconstruction using the program EMAN²². For verification, a second set of paired images was collected and used to carry out an independent reconstruction (see Supplementary Information).

An initial model was generated using 120 particles and the EMAN program 'startort'. This model was used to produce a set of projections with an angular interval of 3°. The images were aligned with these reference projections to determine the origins and orientations, and a three-dimensional reconstruction was calculated. Cycles of angular refinement were performed with an angular interval of 3° until convergence, with a total of 665 particles contributing to the final average map. Full octahedral symmetry was imposed during the reconstruction. To evaluate the quality of the reconstruction, the data were divided into two groups, and two independent reconstructions were calculated. The resolution of the final reconstruction was estimated as 32 Å, where the Fourier shell correlation between the two independent reconstructions started to fall below 0.5. Maps were rendered using the UCSF Chimera package²³.

Received 8 October; accepted 18 December 2003; doi:10.1038/nature02307.

- Seeman, N. C. DNA in a material world. *Nature* **421**, 427–431 (2003).
- Seeman, N. C. DNA nanotechnology: novel DNA constructions. *Annu. Rev. Biophys. Biomol. Struct.* **27**, 225–248 (1998).
- Wilson, D. S. & Szostak, J. W. *In vitro* selection of functional nucleic acids. *Annu. Rev. Biochem.* **68**, 611–647 (1999).
- Chen, J. H. & Seeman, N. C. Synthesis from DNA of a molecule with the connectivity of a cube. *Nature* **350**, 631–633 (1991).
- Zhang, Y. & Seeman, N. C. The construction of a DNA truncated octahedron. *J. Am. Chem. Soc.* **116**, 1661–1669 (1994).
- Dorenbeck, A. *DNA Nanostructures by Self-Assembly of Trisiliconucleotidyls*. Thesis, Univ. Bochum (2000).

- Seeman, N. C. Construction of three-dimensional stick figures from branched DNA. *DNA Cell Biol.* **10**, 475–486 (1991).
- Seeman, N. C. Nucleic-acid junctions and lattices. *J. Theor. Biol.* **99**, 237–247 (1982).
- Seeman, N. C. *De novo* design of sequences for nucleic acid structural engineering. *J. Biomol. Struct. Dyn.* **8**, 573–581 (1990).
- Li, X. J., Yang, X. P., Qi, J. & Seeman, N. C. Antiparallel DNA double crossover molecules as components for nanoconstruction. *J. Am. Chem. Soc.* **118**, 6131–6140 (1996).
- Larson, S. B. *et al.* Double-helical RNA in satellite tobacco mosaic virus. *Nature* **361**, 179–182 (1993).
- Tang, L. *et al.* The structure of Parvovirus reveals a dodecahedral cage of duplex DNA. *Nature Struct. Biol.* **8**, 77–83 (2001).
- Zhang, X., Yan, H., Shen, Z. & Seeman, N. C. Paranemic cohesion of topologically-closed DNA molecules. *J. Am. Chem. Soc.* **124**, 12940–12941 (2002).
- Sa-Ardyen, P., Volododskii, A. V. & Seeman, N. C. The flexibility of DNA double crossover molecules. *Biophys. J.* **84**, 3829–3837 (2003).
- Dubochet, J. *et al.* Cryo-electron microscopy of vitrified specimens. *Q. Rev. Biophys.* **21**, 129–228 (1988).
- Frank, J. Averaging of low exposure electron micrographs of non-periodic objects. *Ultramicroscopy* **1**, 159–162 (1975).
- Frank, J. Classification of macromolecular assemblies studied as 'single particles'. *Q. Rev. Biophys.* **23**, 281–329 (1990).
- Seeman, N. C. & Kallenbach, N. R. Design of immobile nucleic acid junctions. *Biophys. J.* **44**, 201–209 (1983).
- Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31**, 1–10 (2003).
- Stemmer, W. P., Cramer, A., Ha, K. D., Brennan, T. M. & Heyneker, H. L. Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. *Gene* **64**, 49–53 (1995).
- Carragher, B. *et al.* Leginon: an automated system for acquisition of images from vitreous ice specimens. *J. Struct. Biol.* **132**, 33–45 (2000).
- Ludtke, S. J., Baldwin, P. R. & Chiu, W. EMAN: semiautomated software for high-resolution single-particle reconstructions. *J. Struct. Biol.* **128**, 82–97 (1999).
- Huang, C. C., Couch, G. S., Pettersen, E. F. & Ferrin, T. E. Chimera: An extensible molecular modeling application constructed using standard components. *Pacif. Symp. Biocomput.* **1**, 724 (1996).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank F. Guerra for help with EMAN reconstructions and B. Carragher, R. Milligan and C. Potter for advice on reconstructions. This work was supported by the National Aeronautics and Space Administration and The Skaggs Institute for Chemical Biology at The Scripps Research Institute. Some of the work presented here was conducted at the National Resource for Automated Molecular Microscopy, which is supported by the National Institutes of Health through the National Center for Research Resources. W.M.S. is a Fellow supported by the Damon Runyon Cancer Research Foundation. Molecular graphics images were produced using the UCSF Chimera package.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to W.M.S. (wmshih@scripps.edu).

Earthquake nucleation by transient deformations caused by the *M* = 7.9 Denali, Alaska, earthquake

J. Gombert¹, P. Bodin², K. Larson³ & H. Dragert⁴

¹US Geological Survey, Center for Earthquake Research and Information, 3876 Central Ave. Suite 2, and ²Center for Earthquake Research and Information, The University of Memphis, 3876 Central Ave. Suite 1, Memphis, Tennessee 38152-3050, USA

³Department of Aerospace Engineering Sciences, University of Colorado at Boulder, Boulder, Colorado 80309-0429, USA

⁴Geological Survey of Canada, Pacific Geoscience Centre, 9860 West Saanich Road, Sidney, British Columbia, V8L 4B2, Canada

The permanent and dynamic (transient) stress changes inferred to trigger earthquakes are usually orders of magnitude smaller than the stresses relaxed by the earthquakes themselves, implying that triggering occurs on critically stressed faults^{1–4}. Triggered seismicity rate increases may therefore be most likely to occur in areas where loading rates are highest and elevated pore pressures,

perhaps facilitated by high-temperature fluids, reduce frictional stresses and promote failure^{5–7}. Here we show that the 2002 magnitude $M = 7.9$ Denali, Alaska, earthquake triggered widespread seismicity rate increases throughout British Columbia and into the western United States. Dynamic triggering by seismic waves should be enhanced in directions where rupture directivity focuses radiated energy, and we verify this using seismic and new high-sample GPS recordings of the Denali mainshock. These observations are comparable in scale only to the triggering caused by the 1992 $M = 7.4$ Landers, California, earthquake¹, and demonstrate that Landers triggering did not reflect some peculiarity of the region or the earthquake. However, the rate increases triggered by the Denali earthquake occurred in areas not obviously tectonically active, implying that even in areas of low ambient stressing rates, faults may still be critically

stressed and that dynamic triggering may be ubiquitous and unpredictable.

The Denali earthquake ruptured nearly 300 km of fault, propagating unilaterally southeastwards^{8,9} (Fig. 1). The system of faults that broke comprises part of the Pacific–North American plate boundary, which is defined by a series of strike-slip faults that continue through the Alaska Panhandle, the southwestern Yukon, and back into Alaska. Farther south the Juan de Fuca plate subducts obliquely beneath North America. Notably, to first order, the plate-motion-related active deformation is confined to the coastal and offshore regions of Alaska and British Columbia, and the interior of the Cordillera appears relatively inert (Fig. 1)¹⁰.

Seismicity reported in US and Canadian seismicity catalogues was too sparse to reliably detect any rate changes caused by the Denali earthquake, except in a number of densely monitored volcanic and geothermal areas where triggered seismicity was observed at distances as far as 3,660 km (ref. 8). However, networks of broadband seismographs provided data useful for measuring seismicity rate changes at magnitude levels below regional network detection thresholds. We visually examined approximately 16 hours of broadband data centred temporally on the time of the Denali earthquake at all broadband stations in western Canada and most in the western US. Networks that supplied data include the Canadian National Seismic, the Global Seismic, Pacific Northwest Seismograph, Berkeley Digital Seismic, ANZA Regional and the US National Networks; the Geologic Survey of Canada and the Incorporated Research Institutions for Seismology Data Management System disseminated the data. Whereas the ambient noise and surface waves from the Denali mainshock dominate these data, with frequencies in the 0.2 to 0.02-Hz range, high-pass filtering reveals the higher-frequency signals radiated by small earthquakes local to the recording site (Fig. 2). Unfortunately data retrieval and analysis cannot be automated reliably and are time-consuming; examination of longer time windows is part of ongoing work looking at the duration of the rate increases and their longer-term behaviour.

Although locations cannot be determined from a single recording, the time difference between arriving P and S waves provides estimates of the distance from the station to each earthquake. We only catalogued events with S–P differences of <30 s, corresponding approximately to <240 km, with most events having considerably shorter times and distances. At numerous sites the number of local

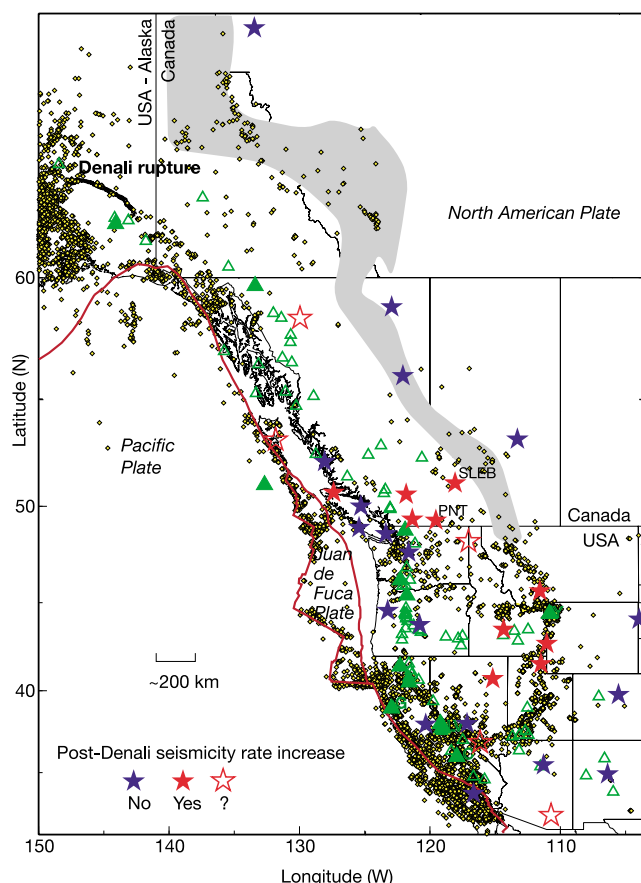


Figure 1 Map of western North America showing the Denali earthquake faults, tectonic features and sites of seismicity rate change observations. Shown are the surface trace of the Denali earthquake rupture (black curve), epicentres of pre-Denali earthquakes with magnitudes >2.0 (yellow dots) for a period of five years before the Denali earthquake (3 November 1997–2002), sites of volcanic activity of the Holocene epoch (open green triangles) and historic age (solid green triangles) (see <http://www.volcano.si.edu/world/suminfo.cfm>), and broadband seismograph station sites where seismicity rates were examined before and after the Denali earthquake (stars). Star colours indicate whether the seismicity rate increased at the time of the Denali earthquake. Red lines outline plate boundaries. The Pacific–North American boundaries in Canada southward are transform fault zones, the Juan de Fuca–Pacific boundary ridges and transform faults. The Juan de Fuca plate subducts beneath North America, resulting in volcanic chains. The grey band in Canada outlines the 'morphogeologic' transition between deformed strata on the west and flat-lying undeformed strata to the east, or the boundary with stable North America¹⁹. Geothermal areas where the Denali earthquake triggered seismicity rate increases that are documented elsewhere⁸ are not shown.

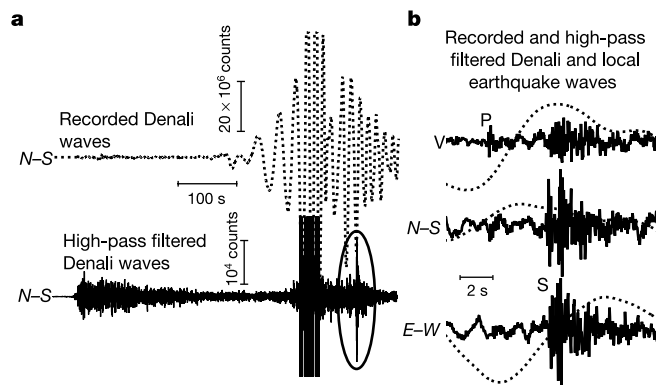


Figure 2 Local earthquake signals revealed in filtered seismograms. **a**, North–south (N–S) component of the Denali mainshock ground velocities (in units of digital counts; dotted trace) at station PNT (Fig. 1) recorded by a broadband seismograph. After filtering out frequencies below ~ 2 Hz (solid trace) a local earthquake signal (circled) becomes apparent. **b**, Expanded view of the recorded and filtered waveforms within the circled time window on the left, for all three components of ground velocity (V is vertical, $E-W$ is east–west). The arrival times of the P and S waves are clearly observable in the filtered waveforms.

events was significantly greater in the time period after the Denali earthquake than before; at some sites the rate was comparable in the before and after periods, and at others there were only a few local events in the post-Denali period—also indicative of no rate change (Fig. 3). The probability that the latter effect is due instead to a very low background rate (requiring a longer time to see a change) seems low, because similar observations were also made at sites with relatively high background rates. At all triggered sites the first clearly identifiable local, presumably triggered, earthquake occurred within minutes following the arrival of the Denali P wave and after the largest-amplitude surface waves within slower, smaller surface and coda waves (that is, while the Denali signal was above the background noise level). Although seismicity rates fluctuate and may increase in the absence of any obvious external cause, the coincidence of rate-increase onset times near the Denali signal suggests that most, if not all, were not chance occurrences. We also note that in no case was the rate increase delayed from the mainshock signal according to a simple statistical criterion¹, consistent with most observations and theoretical models of aftershocks at near-source distances.

Despite the uneven sampling, the distribution of sites where the seismicity rate increased following the Denali earthquake clearly was not random and extended to distances of 3,385 km from the Denali epicentre, far beyond what is traditionally considered the aftershock region or where permanent stress changes have significant magnitude^{1–3} (Fig. 3). As Fig. 1 shows,

the distribution of Denali-triggered sites does not correlate with the regions of highest tectonic loading rates, which should lie close to the plate boundaries and be evident as areas with the greatest ambient seismicity rates^{11,12}. Moreover, the locations of triggered seismicity rate increases show no clear correlation with active volcanic or geothermal sites. Particularly striking are the observations in south-central British Columbia. These include triggering at station SLEB, which lies in the foothills of the Canadian Cordillera, almost at the boundary of what is considered to be ancient stable North America, on the basis of geologic and morphologic studies¹⁹. Just slightly to the west, at station PNT, geodetic studies show that active deformation is ‘suspected’ but occurs at significantly lower rates than within what is considered the active plate boundary zone¹¹—where almost no triggering occurred. A recent analysis of geophysical data in western Canada shows the area of triggered seismicity to be in a transitional zone between old, cold, strong crust to the east and younger, warmer and weaker crust to the west¹³. Our results are consistent with the idea that critically stressed faults may exist nearly anywhere, as inferred from human-induced seismicity, apparently triggered historic earthquakes, and deep drilling studies in intra-plate settings^{14–16}.

Although there are large gaps in data coverage, Denali triggering plausibly occurs within a continuous band extending from the rupture as far south as northern Utah and Nevada. This band lies approximately on-strike with the Denali fault rupture, which also marks the direction of maximum Love-wave radiation that

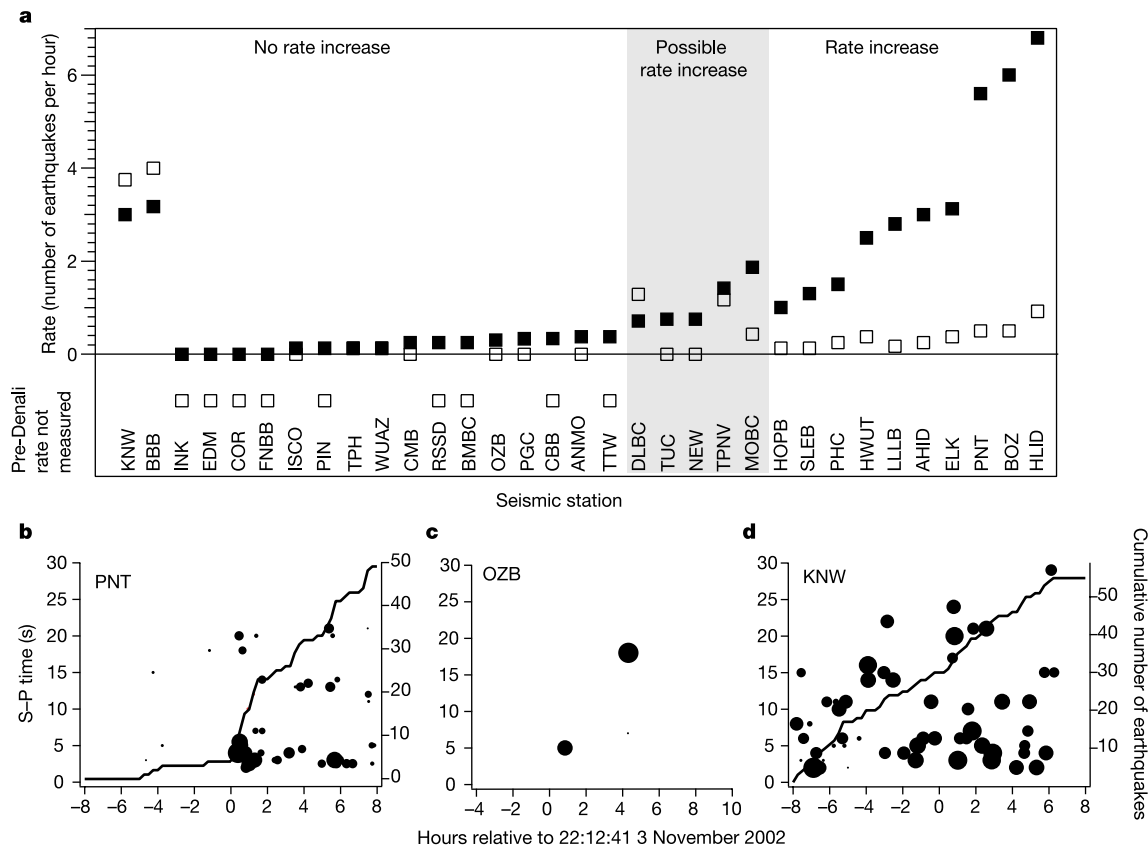


Figure 3 Seismicity rate change observations. **a**, Rates estimated as the number of earthquakes per hour before (open squares) and after (filled squares) the Denali earthquake at each seismic station examined (labelled). At stations where the pre-Denali rate is graphed below zero, no pre-Denali data were examined because the post-Denali rate was insignificant. At the stations where a rate increase was deemed possible but ambiguous (within grey band), the rate increases for a restricted distance (S–P time) range and/or for events bigger than some size limit. **b–d**, Examples of S–P times (left axis, dot

size proportional to $\log[\text{peak signal amplitude}]$) of filtered earthquake signals before and after Denali waves arrive. The cumulative pre- and post-Denali earthquake counts (lines) show the rate change or lack of change more clearly. At some sites a post-Denali rate increase is obvious (PNT, Fig. 1), at some the rate does not change (station KNW in southern California), and at others the low post-Denali count implies that either the pre-Denali rate is too low to measure a rate change or that there is none (station OZB on Vancouver Island).

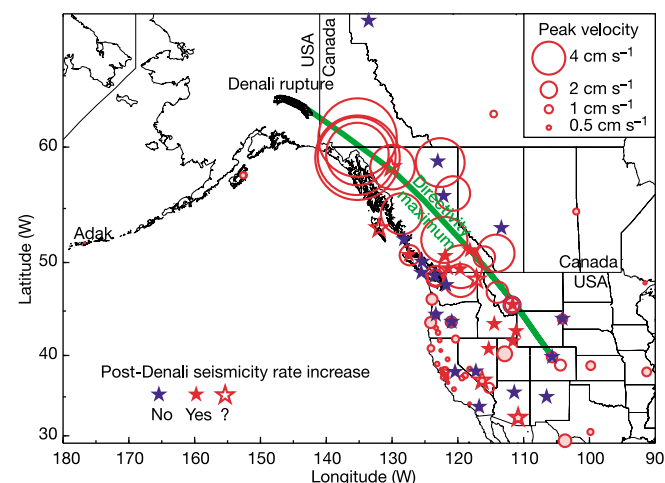


Figure 4 Map of the distributions of Denali-related seismicity rate changes and measured peak Denali seismic ground velocities. The latter are proportional to dynamic strains at these distances²⁰ and were recorded at broadband seismic stations (shaded circles) and 1-Hz GPS receivers (open circles). Circle centres show locations of recording sites with radii proportional to the measured peak velocity (scale in inset). All measurements in British Columbia and four in the US are from 1-Hz GPS data: note the good agreement with the nearest seismic measurement. The velocities decrease away from the direction of theoretically expected maximum seismic radiation (green segment of a great circle). We label the seismic station at Adak because the measured velocities are too small to be visible. Notably, the sites of triggered rate increases lie roughly within the same azimuthal span as is defined by the maximum measured velocities and theoretically expected dynamic deformations.

dominates the wavefield from a strike-slip earthquake mechanism like the Denali event^{8,9}. Moreover, the propagation of the rupture from north to south should focus seismic wave energy along-strike southward of the rupture owing to the focusing phenomena of directivity. We verify these theoretical predictions by mapping the maximum Denali mainshock amplitudes recorded at broadband seismic stations and new permanent GPS receivers that sample at 1 Hz (ref. 17). Canadian GPS data were provided by the Geodetic Survey Division, Geomatics Canada, and Base Mapping and Geomatic Services, British Columbia Ministry of Sustainable Resource Management. Other contributing geodetic networks include the International GPS Service, the Continuously Operating Reference Stations and the University NAVstar Consortium and Aeromap Services. This new application of GPS technology for measuring seismic waves is essential to defining the pattern of seismic radiation, because the Denali seismic waves saturated all the broadband seismographs within British Columbia and most within the most northwestern US.

The combined GPS and seismic data clearly show a focusing of seismic wave energy, or dynamic deformations, consistent with the theoretical expectations (Fig. 4). The coincidence of this focusing with the band of triggered seismicity rate increases strongly suggests to us that the dynamic deformations were the triggering agent. However, while this correlation is clearly suggestive, there is not a one-to-one correspondence between the peak amplitude of the dynamic deformation and the change in seismicity rate. This might imply a triggering threshold that depends on tectonic environment¹⁸. Alternatively or additionally, it might suggest that some other characteristic of the dynamic deformation is important in nucleating rupture (for example, duration or frequency of oscillation). The unprecedented abundance of direct measurements of the dynamic deformation field, albeit at the surface, provides new opportunities to investigate these possibilities. □

Received 14 August 2003; accepted 13 January 2004; doi:10.1038/nature02335.

- Hill, D. P. *et al.* Remote seismicity triggered by the M7.5 Landers, California earthquake of June 28, 1992. *Science* **260**, 1617–1623 (1993).
- Gomberg, J., Beeler, N. & Blanpied, M. On rate-state and Coulomb failure models. *J. Geophys. Res.* **105**, 7857–7872 (2000).
- Huc, M. & Main, I. G. Anomalous stress diffusion in earthquake triggering: Correlation length, time dependence, and directionality. *J. Geophys. Res.* **108**(B7), doi:10.1029/2001JB001645 (2003).
- Belardinelli, M. E., Bizzarri, A. & Cocco, M. Earthquake triggering by static and dynamic stress changes. *J. Geophys. Res.* **108**(B3), doi:10.1029/2002JB001779 (2003).
- Cocco, M. & Rice, J. R. Pore pressure and poroelasticity effects in Coulomb stress analysis of earthquake interactions. *J. Geophys. Res.* **107**(B2), doi:10.1029/2000JB000138 (2002).
- Streit, J. E. & Cox, S. F. Fluid pressures at hypocenters of moderate to large earthquakes. *J. Geophys. Res.* **106**, 2235–2243 (2001).
- Sturtevant, B., Kanamori, H. & Brodsky, E. E. Seismic triggering by rectified diffusion in geothermal systems. *J. Geophys. Res.* **101**, 25269–25282 (1996).
- Eberhart-Phillips, D. *et al.* The 2002 Denali Fault Earthquake, Alaska: A large magnitude, slip-partitioned event. *Science* **300**, 1113–1118 (2003).
- Ozcar, A. A., Beck, S. & Christiansen, D. H. Source process of the 3 November 2002 Denali fault earthquake (central Alaska) from teleseismic observations. *Geophys. Res. Lett.* **30**, doi:10.1029/2003GL017272 (2003).
- Riddiough, R. P. & Hyndman, R. D. in *Geology of the Cordilleran Orogen in Canada* (eds Gabrielse, H. & Yorath, C. J.) Ch. 13, no. 4, 435–455 (Geologic Survey of Canada, Geology of Canada, Ottawa, 1991).
- Miller, M. M. *et al.* GPS-determination of along-strike variation in Cascadia margin kinematics: Implications for relative plate motion, subduction zone coupling, and permanent deformation. *Tectonics* **20**, 161–176 (2001).
- Anderson, J. G. *et al.* Seismicity in the western Great Basin apparently triggered by the Landers, California, earthquake 28 June 1992. *Bull. Seismol. Soc. Am.* **84**, 863–891 (1994).
- Flück, P., Hyndman, R. D. & Lowe, C. Effective elastic thickness T_e of the lithosphere in western Canada. *J. Geophys. Res.* **108**(B9), doi:10.1029/2002JB002201 (2003).
- Seeber, L., Armbruster, J. G., Kim, W.-Y. & Barstow, N. The 1994 Cacoosing Valley earthquakes near Reading, Pennsylvania: A shallow rupture triggered by quarry unloading. *J. Geophys. Res.* **103**, 24505–24521 (1998).
- Hough, S. E., Seeber, L. & Armbruster, J. G. Intraplate triggered earthquakes: Observations and interpretation. *Bull. Seismol. Soc. Am.* **93**, 2212–2221 (2003).
- Townend, J. & Zoback, M. D. How faulting keeps the crust strong. *Geology* **28**, 399–402 (2000).
- Larson, K., Bodin, P. & Gomberg, J. Using 1 Hz GPS data to measure deformations caused by the Denali fault earthquake. *Science* **300**, 1421–1424 (2003).
- Gomberg, J., Reasenberg, P., Bodin, P. & Harris, R. Earthquake triggering by transient seismic waves following the Landers and Hector Mine, California earthquakes. *Nature* **411**, 462–466 (2001).
- Gabrielse, H., Monger, J. W. H., Wheeler, J. O. & Yorath, C. J. in *Geology of the Cordilleran Orogen in Canada* (eds Gabrielse, H. & Yorath, C. J.) Ch. 2, no. 4, 13–59 (Geologic Survey of Canada, Geology of Canada, Ottawa, 1991).
- Gomberg, J. & Agnew, D. C. The accuracy of seismic estimates of dynamic strains from Pinon Flat Observatory, California, strainmeter and seismograph data. *Bull. Seismol. Soc. Am.* **86**, 212–220 (1996).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.G. (gomberg@usgs.gov).

Ramp initiation in a thrust wedge

John Panian & David Wiltschko

Department of Geology and Geophysics and Center for Tectonophysics, Texas A&M University, College Station, Texas 77843-3115, USA

Collisional mountain belts are characterized by fold and thrust belts that grow through sequential stacking of thrust sheets from the interior (hinterland) to the exterior (foreland) of the mountain belt^{1–5}. Each of these sheets rides on a fault that cuts up through the stratigraphic section on inclined ramps that join a flat basal fault at depth. Although this stair-step or ramp-flat geometry is well known, there is no consensus on why a particular ramp forms where it does. Perturbations in fault shape^{6,7}, stratigraphy^{8,9}, fluid pressure^{10,11}, folding^{2,12}, and surface slope^{13,14} have all been suggested as possible mechanisms. Here we show that such pre-existing inhomogeneities, though feasible causes, are not required. Our computer simulations show that a broad foreland-dipping plastic strain band forms at the surface near the topographic inflection produced by the previous ramp. This

strain band then migrates towards the rigid base, where the plastic strain is preferentially concentrated in a thrust ramp. Subsequent ramps develop toward the foreland in a similar fashion. Syntectonic erosion and deposition may strongly control the location of thrust ramps by enhancing or removing the surface point of initiation.

Restored cross-sections from diverse fold and thrust belts show that thrust ramp spacing is a function of thrust-sheet thickness^{15–21} (Fig. 1). Specifically, ramp spacing decreases toward the foreland, in the same direction that the sedimentary package caught up in the thrusting thins. This observation suggests both that the mechanism for initiating thrust ramps is fundamental to the mechanics of fold and thrust belts and, while feasible and well-documented in some cases²², local perturbations are not necessary to form thrust ramps.

We have conducted computer experiments to test whether ramp spacing can be explained without calling on local inhomogeneities. Our models use a modified Drucker–Prager constitutive relation that is a pressure-dependent, elastic–perfectly plastic rheology; see Methods. We incorporate an initial wedge-shape geometry that replicates the thinning of the sedimentary package toward the foreland^{23–25}. The geometry of our finite-element model is 100 km in length and has a basal slope equal to the surface slope of 2°. The wedge lies upon a basal, rigid surface that represents basement rocks. The basal boundary condition is Coulomb frictional sliding. The rear wedge boundary is advanced in 10-m increments.

To characterize the initiation of a new thrust ramp, we begin at the point where the first backthrust–forethrust pair has formed at the rear of the wedge at 2,500 m of backstop displacement (Fig. 2a). It is not important how the first ramp forms. For example, we have run models where the basal friction abruptly changes. The point of change in friction is a ramp nucleation point. However, subsequent ramps form in the same manner as we show here. We have also run models with a basal perturbation in the form of a small step in the rigid base. In these models ramp development is essentially the same. In natural fold and thrust belts, the first ramp ultimately is the plate boundary.

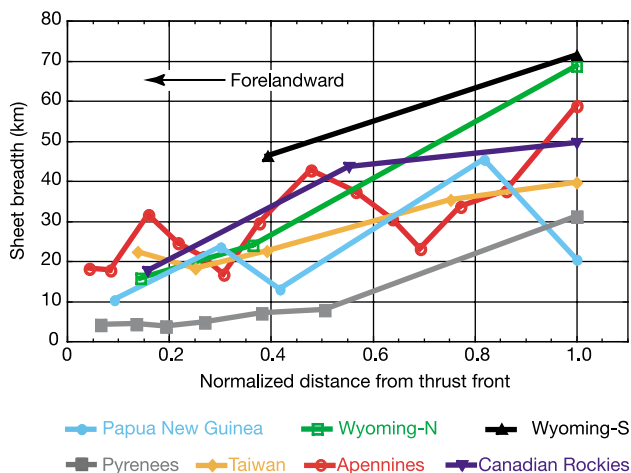


Figure 1 Thrust sheet breadth versus restored distance from the thrust front. The restored distance is measured to the most hinterlandward ramp and normalized to the total restored length of thrust sheets measured. Only thrust sheets having preserved hanging-wall cutoffs are included. Some grouping of minor splays was done. Sources were the Apennines¹⁵, northern Wyoming¹⁶, Taiwan¹⁷, southern Wyoming¹⁸, the Canadian Rockies¹⁹, Papua New Guinea²⁰ and the Pyrenees²¹. Owing to the incomplete nature of subsurface data, the variability in both the Apennine and Papua New Guinea cases could be decreased by slight reinterpretation of the seismic data. Nevertheless, as drawn, ramp spacing decreases toward the thrust front or foreland, in the same direction that the stratigraphic thickness decreases.

After the initial ramp is formed (Fig. 2a), applying the next 10 m of displacement on the left boundary or backstop causes the structures to slide along the base. The basal displacement quickly tapers off to about 2 m displacement approximately 18 km from the origin and trends towards zero forelandward with a subtle change in slope near 35 km (Fig. 2b). This first thrust sheet is moving up its ramp at ~35° (Fig. 2c).

The next ramp to form, in terms of plastic strain concentration, is initiated at the surface inflection of the previous existing ramp, and propagates in a broad band toward the base of the wedge. At 2,820 m of rear displacement, the band of plastic strain meets the base and begins to develop towards the foreland, initiating a new thrust ramp (Fig. 3a). The precursors for ramp initiation are: (1) increased basal displacement out to 35 km from the origin for this particular model (Fig. 3b) and (2) change in uplift path of the surface from nearly parallel to the thrust-sheet base to being inclined to it (Fig. 3c).

At 3,070 m of rearward displacement, a new thrust ramp is fully formed ~37 km from the origin (Fig. 4a). The forethrust clearly predominated over the backthrust in terms of plastic strain intensity. The basal displacement along the new thrust sheet has increased twofold (Fig. 4b) and the uplift path of the surface of the new thrust sheet approaches that of the first.

The pattern of displacement of the model surface illustrates the evolution of ramps. Figure 5a shows the difference in surficial displacement vectors between 2,500 and 2,820 m of backstop displacement. The onset of formation of the second ramp is clearly indicated 30 km from the backstop. Displacement vectors for material on the first ramp shallow. Simultaneously, the second ramp dominates abruptly as shown in the 2,820–3,070 m backstop displacement increment (Fig. 5b). The material on the first ramp has acquired a significant component of horizontal motion owing to displacement on the second ramp. This trend continues for subsequent ramps: motion on successive ramps adds a component

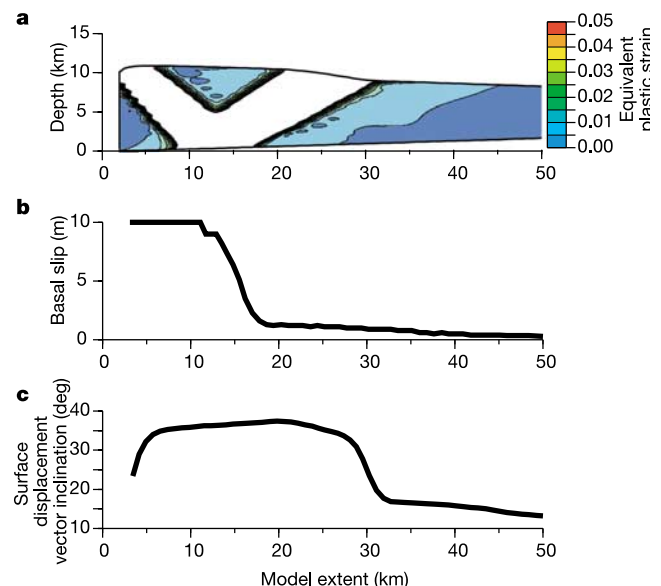


Figure 2 Thrust wedge after 2,500 m of backstop (left boundary) displacement.

a, Equivalent plastic strains. A subtle increase in basal displacement in area **b** and the beginning of surface uplift. **b**, Basal displacement due to just the last 10-m increment of backstop displacement, that is, 2,490 to 2,500 m of backstop displacement. **c**, Inclination of displacement vectors on the surface of the model. The backthrust is subtly beginning, as indicated by the deflection in the 0.005 equivalent plastic strain contour at about 35 km in **a**, and in the same broad area in **c**. Compare Figs 3 and 4. Strains above 0.050 in **a** have been blocked out to emphasize deformation in the footwall. Although the model extends horizontally to 100 km, only the first 50 km are shown.

of horizontal motion to more interior ones and therefore their net motion relative to an external fixed reference frame appears to be shallower.

The pattern of surficial displacement vectors may indicate the location of a new thrust ramp prior to significant development and movement along that ramp. Geodetic measurements in active fold and thrust belts may be able to record this pattern. Inhomogeneities of the earthquake cycle may obscure the pattern for slowly moving faults.

The initiation of a thrust ramp is analogous to perturbation models for thrust duplex initiation^{26,27}. Duplexes are found where ramps are constrained between quasi-parallel upper and lower thrust faults. The ramps link one with the other. In our wedge models there is no upper or roof thrust. Ramps initiate at the surface where an inflection has formed at the toe of a pre-existing ramp. Once this band reflects from the rigid base at the location of the basal or floor thrust, a conjugate, backthrust–forethrust set of plastic bands quickly localizes, with the forethrust band preferentially accumulating more plastic strain with backstop displacement. By analogy with duplexes, our perturbation would be the topographic inflection at the surface produced by a pre-existing thrust. This emphasizes the conclusion that perturbations are still required to form ramps. However, the perturbations we propose do not exist until the fold and thrust belt begin to develop. In this way our results are similar to those of ref. 12, which showed that thrust sheet spacing may be explained by folding. Specifically, the decrease in thrust spacing toward the foreland tracks wedge thinning because fold wavelength is strongly a function of layer thickness. Folds forming in front of the last ramp would localize thrust ramps at successively closer spacings toward the foreland.

Once a new thrust localizes, the previous ramp region behind the new thrust is displaced along the base with relatively little additional plastic straining. Most of the shortening that does take place behind the ramp causes uplift and shearing along the ramp. This surface displacement should appear in geodetically active fold and thrust belts as bulk horizontal translation with some uplift. Almost all

plastic strain is located in the ramp (forethrust) regions and in their conjugate backthrusts. Immediately preceding localization of a new ramp, broad bands of plastic strain develop from the surface of the wedge towards the base. As the bands meet the base, they reflect upwards towards the surface of the wedge at an angle. The bands quickly, in terms of backstop displacement steps, become narrower, forming a backthrust–forethrust pair. With increasing displacement, the forethrust intensifies in plastic strain, while the backthrust lessens; successive thrust ramps form in a similar way within a thrust

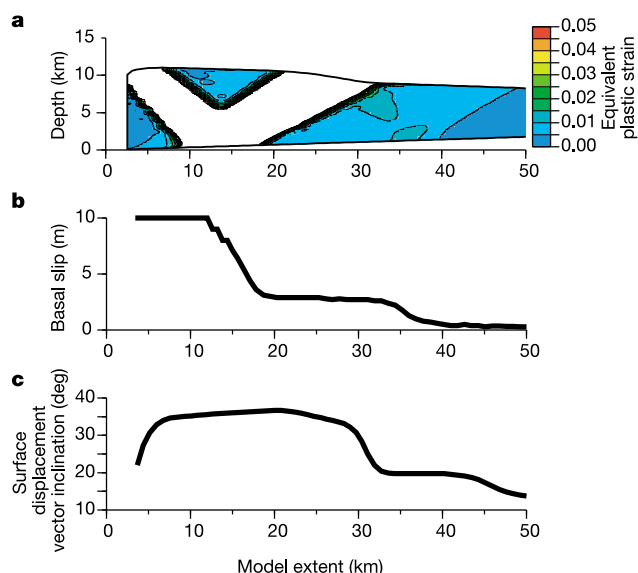


Figure 3 Thrust wedge deformation after 2,820 m of backstop displacement. **a**, Equivalent plastic strains. **b**, Basal displacement due to last 10-m increment of backstop displacement. **c**, Inclination of displacement vectors on the surface of the model. The conjugate shear band is now well developed, as can be seen in the 0.010 strain contour (**a**), increased basal slip behind the forming ramp (**b**) and an increase in inclination of surface displacement (**c**).

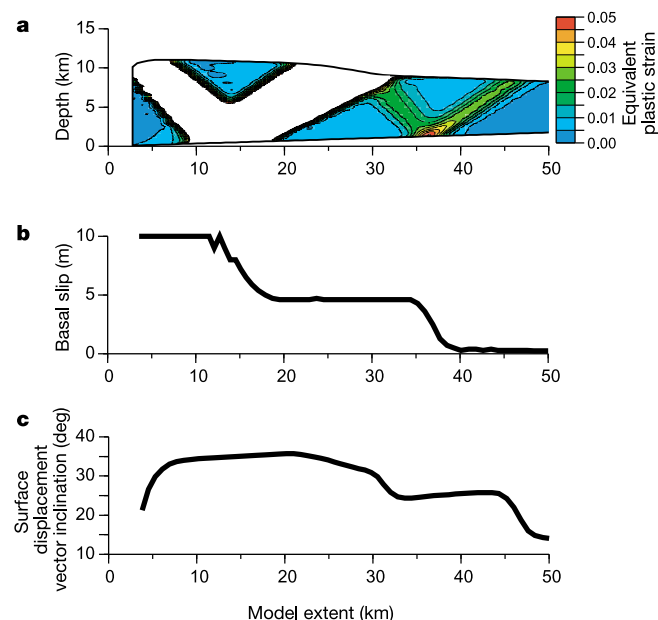


Figure 4 Thrust wedge deformation after 3,070 m of backstop displacement. **a**, Equivalent plastic strains. **b**, Basal displacement due to last 10-m increment of backstop displacement. **c**, Inclination of displacement vectors on the surface of the model. The forward ramp is now well developed with strains up to 0.050 at the base (**a**), basal slip nearly half that of the backstop (**b**) and an 30° inclination of surface displacement, that is, a typical ramp angle (**c**).

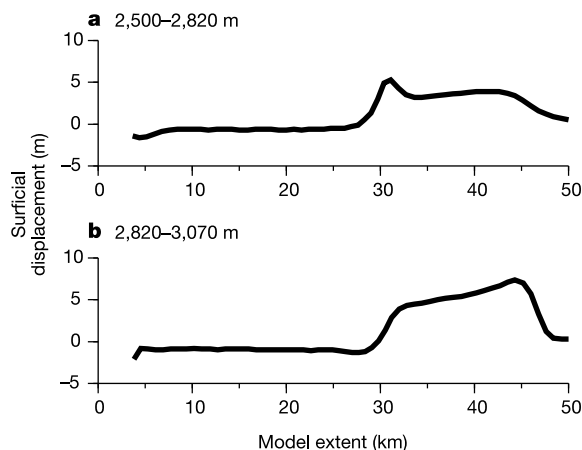


Figure 5 Magnitude of model surface displacement accumulated during two intervals of backstop displacement. **a**, Surface displacement magnitude accumulated from 2,500 to 2,820 m of backstop displacement. **b**, Surface displacement magnitude accumulated from 2,820 to 3,070 m of backstop displacement. Note the marked increase in surface displacement from ramp initiation (**a**) to the production of a well-developed ramp (**b**).

wedge. We propose that the strain intensification of each forethrust is a product of the build-up of the surface slope behind the forethrust and the tapered geometry of the wedge. □

Methods

We used the commercial finite-element software package, ABAQUS, to calculate the plastic strains and displacements in our models²⁸. The yield function, F , is defined as:

$$F = aq^b - p - p_t \quad (1)$$

where p is the mean stress, q , the equivalent Mises stress, is defined as:

$$q = \sqrt{\frac{3}{2} \sigma'_{ij} \sigma'_{ij}} = \sqrt{3} J_2 \quad (2)$$

J_2 is the second variant of deviatoric stress, where $\sigma'_{ij} = \sigma_{ij} - p\delta_{ij}$ is the deviatoric stress tensor, p_t represents the mean pressure tensile strength of the material. a and b in equation (1) are material parameters, independent of plastic deformation, that define the concavity of the yield surface in stress space. The three material parameters defined above (a , b , and p_t) are determined by a least-squares fit to triaxial test data. Data for our models were limited to those studies done on sedimentary rock under at least three different confining pressures no larger than about 250 MPa, or approximately 10 km depth. We used yield stress versus confining pressure values for a sandstone²⁹ that were about average for all the sedimentary rocks for which data exist. The material parameters in equation (1) are $a \approx b \approx 1$ and $p_t \approx 0$, which is identical for most of the yield stress data used in our models. The equivalent plastic strain measure^{28,30} is defined as:

$$\varepsilon^p = \int \sqrt{\frac{\sigma_{ij} d\epsilon_{ij}^p}{p}} \quad (3)$$

Received 10 June 2003; accepted 13 January 2004; doi:10.1038/nature02334.

- Bally, A. W., Gordy, P. L. & Stewart, G. A. Structure, seismic data, and orogenic evolution of the southern Canadian Rocky Mountains. *Can. J. Petrol. Geol.* **14**, 337–381 (1966).
- Dahlstrom, C. D. A. Structural geology in the eastern margin of the Canadian Rocky Mountains. *Bull. Can. Petrol. Geol.* **18**, 332–406 (1970).
- Armstrong, F. C. & Oriel, S. S. Tectonic development of Idaho-Wyoming thrust belt. *Am. Assoc. Petrol. Geol. Bull.* **71**, 1847–1866 (1965).
- Wiltschko, D. V. & Dorr, J. A. Timing and deformation in overthrust belts and foreland of Idaho, Wyoming, and Utah. *Am. Assoc. Petrol. Geol. Bull.* **67**, 1304–1322 (1993).
- Jordan, T. E., Allmendinger, R. W., Damanti, J. F. & Drake, R. E. Chronology of motion in a complete thrust belt; the Precordillera, 30–31° S, Andes Mountains. *J. Geol.* **101**, 135–156 (1993).
- Wiltschko, D. V. & Eastman, D. B. Role of basement warps and faults in localizing thrust-fault ramps. *Geol. Soc. Am. Mem.* **158**, 177–190 (1982).
- Knappe, R. J. Footwall geometry and the rheology of thrust sheets. *J. Struct. Geol.* **7**, 1–10 (1985).
- Bombalakis, E. G. Thrust-fault mechanics and origin of a frontal ramp. *J. Struct. Geol.* **8**, 281–290 (1986).
- Platt, J. P. The mechanics of frontal imbrication, a first-order analysis. *Geologische Rundschau* **77**, 577–589 (1986).
- Davis, D., Suppe, J. & Dahlen, F. A. Mechanics of fold-and-thrust belts and accretionary wedges. *J. Geophys. Res.* **88**, 1153–1172 (1983).
- Cello, G. & Nurr, A. Emplacement of foreland thrust systems. *Tectonics* **7**, 261–272 (1988).
- Goff, D. F., Wiltschko, D. V. & Fletcher, R. C. Decollement folding as a mechanism for thrust-ramp spacing. *J. Geophys. Res.* **101**, 11341–11352 (1996).
- Panian, J. & Pilant, W. L. A possible explanation for foreland thrust propagation. *J. Geophys. Res.* **86**, 8607–8615 (1990).
- Goff, D. F. & Wiltschko, D. V. Stresses beneath a ramping thrust sheet. *J. Struct. Geol.* **14**, 437–449 (1992).
- Bally, A. W., Burbil, L., Cooper, C. & Ghelardoni, R. Balanced sections and seismic reflection profiles across the central Apennines. *Mem. Soc. Geol. It.* **35**, 257–310 (1986).
- Royce, F. Jr, Warner, M. A. & Reese, D. L. *Deep Drilling Frontiers of the Central Rocky Mountains* 41–54 (Rocky Mountain, Assoc. Geol. 1975).
- Suppe, J. A retrodeformable cross section of northern Taiwan. *Proc. Geol. Soc. China* **23**, 46–55 (1980).
- Dixon, J. S. Regional structural synthesis, Wyoming salient of western overthrust belt. *Am. Assoc. Petrol. Geol. Bull.* **66**, 1560–1580 (1982).
- Price, R. A. & Mountjoy, E. W. Geologic structure of the Canadian Rocky Mountains between Bow and Athabasca rivers—a progress report. Structure of the southern Canadian Cordillera. *Geol. Assoc. Can. Spec. Pap.* **6**, 7–25 (1970).
- Hill, K. C. Structure of the Papuan fold belt, Papua New Guinea. *Am. Assoc. Petrol. Geol. Bull.* **75**, 857–872 (1991).
- Williams, G. D. Thrust tectonics in the south central Pyrenees. *J. Struct. Geol.* **7**, 11–17 (1985).
- Kraig, D. H., Wiltschko, D. V. & Spang, J. H. Interaction of basement uplift and thin skinned-thrusting, Moxa arch and the Western Overthrust Belt, Wyoming: A hypothesis. *Am. Assoc. Petrol. Geol. Bull.* **99**, 654–662 (1987).
- Elliot, D. The motion of thrust sheets. *J. Geophys. Res.* **81**, 949–963 (1976).
- Chapple, W. M. Mechanics of thin-skinned fold-and-thrust belts. *Geol. Soc. Am. Bull.* **89**, 1189–1198 (1978).
- Makel, G. & Walters, J. Finite-element analysis of thrust tectonics: computer simulation of detachment phase and development of thrust faults. *Tectonophysics* **226**, 167–185 (1993).
- Mandl, G. & Shippam, G. K. Mechanical model of thrust sheet gliding and imbrication. In *Thrust and Nappe Tectonics* 79–98 (Geol. Soc. Lond. Spec. Publ., 1981).
- Davies, R. K. & Fletcher, R. C. Shear bands in a plastic layer at yield under combined shortening and shear: A model for the fault array in a duplex. In *Deformation Mechanisms, Rheology and Tectonics*, 123–131 (Geol. Soc. Lond. Spec. Publ. 54, 1990).

- Hibbitt, H. D., Karlson, B. I. & Sorenson, E. P. *ABAQUS Theory Manual* (Hibbitt, Karlson & Sorenson, Pawtucket, RI, 1997).
- Handin, J. & Hager, R. V. Experimental deformation of sedimentary rocks under confining pressure: Tests at room temperature on dry samples. *Am. Assoc. Petrol. Geol. Bull.* **41**, 1–50 (1957).
- Malvern, L. E. *Introduction to the Mechanics of a Continuous Media* (Prentice-Hall, Englewood Cliffs, NJ, 1969).

Acknowledgements The NSF Tectonics programme supported this work. We wish to thank B. Johnson and J. Melosh for discussions.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.P. (jpanian@neo.tamu.edu).

New light shed on the oldest insect

Michael S. Engel¹ & David A. Grimaldi²

¹Division of Entomology, Natural History Museum, and Department of Ecology & Evolutionary Biology, Snow Hall, 1460 Jayhawk Blvd, University of Kansas, Lawrence, Kansas 66045-7523, USA

²Division of Invertebrate Zoology, American Museum of Natural History, Central Park West at 79th Street, New York, New York 10024-5192, USA

Insects are the most diverse lineage of all life in numbers of species, and ecologically they dominate terrestrial ecosystems. However, how and when this immense radiation of animals originated is unclear. Only a few fossils provide insight into the earliest stages of insect evolution, and among them are specimens in chert from Rhynie, Scotland's Old Red Sandstone (Pragian; about 396–407 million years ago¹), which is only slightly younger than formations harbouring the earliest terrestrial faunas. The most well-known animal from Rhynie is the springtail *Rhyniella praecursor* (Entognatha; Collembola), long considered to be the oldest hexapod^{2,3}. For true insects (Ectognatha), the oldest records are two apparent wingless insects from later in the Devonian period of North America^{4,5}. Here we show, however, that a fragmentary fossil from Rhynie, *Rhyniognatha hirsti*, is not only the earliest true insect but may be relatively derived within basal Ectognatha. In fact, *Rhyniognatha* has derived characters shared with winged insects, suggesting that the origin of wings may have been earlier than previously believed. Regardless, *Rhyniognatha* indicates that insects originated in the Silurian period and were members of some of the earliest terrestrial faunas.

Traditional interpretation of palaeontological evidence indicates that insects had their beginnings in a group of small, scurrying hexapods in the earliest Devonian period. The most basal insects are bristletails and silverfish, wingless scavengers with ectognathous mandibles and elongate, well-developed ovipositors^{6,7}. Until this study, true insects had not been known conclusively before the Emsian and consisted of only two records from the Devonian period. Cuticular fragments attributable to a wingless insect (either Archaeognatha or Zygentoma) were discovered in compressed shales of the Late Devonian of Gilboa, New York (approximately 379 million years (Myr) ago)⁵. Unfortunately, these isolated sclerites cannot be assigned definitively to any particular order of basal insects, greatly limiting their utility for understanding early insect diversification. Slightly older are the compressed shales of Gaspé Bay in Quebec, Canada (about 390 Myr), from which was recovered a bristletail (Archaeognatha)⁴. Curiously, the Gaspé fossil is not compressed unlike most digested fragments from such shales, leading some authors to contend that it is a recent contaminant⁸. A critical re-study of the Gaspé material is necessary.

At the time that *Rhyniella* was discovered and recognized as a springtail, a pair of well-preserved mandibles was identified among other sclerotized debris in a separate piece of chert. The mandibles were later studied by Tillyard who described them as *Rhyniognatha hirsti*⁹. Tillyard noted that *Rhyniognatha* was insect-like but he could not determine its placement within Hexapoda. Later authors also indicated the “suggestive” appearance of the mandibles but no detailed re-description or study was undertaken and conclusive placement of the fossil was left open; indeed, no subsequent authors re-examined the fossil, simply reiterating Tillyard’s assertion⁶. Thus, the Gaspé bristletail was reported as the earliest insect⁴ and *Rhyniognatha* was relegated to fragments of indeterminate origin¹⁰. Indeed, *Rhyniognatha* was even excluded from the most comprehensive account of hexapod fossils¹¹. Again without ever observing the original material, a recent account of the history of insects dismissed *Rhyniognatha* as either being a contaminant or representing some other arthropod¹². The authenticity of the material from Rhynie has already been elaborated elsewhere¹³ and, as we discuss below, the mandibles cannot be assigned to non-hexapod, mandibulate lineages.

Our study of the unique *Rhyniognatha* specimen in the Natural History Museum, London (NHM-Palaeontology, In. 38234), using

compound microscopy has revealed an anterior mandibular articulation, forming a socket (that is, an anterior acetabulum) on the inner angle of the mandibles, demonstrating that they are dicondylar (that is, with two points of articulation and restricted to a single plane of motion) (Fig. 1). *Rhyniognatha* is not transversely oriented or multiarticulated as in Crustacea, but among myriapods might most easily be confused with the millipede gnathal lobe. The millipede mandible is a robust, multiarticulated structure with a broad, sclerotized base to which the toothed gnathal lobe articulates. The gnathal lobe has a large, basal rasping surface, which is not present in *Rhyniognatha*. Furthermore, no evidence of additional, articulated mandibular sclerites exists in *Rhyniognatha* (Fig. 1). Such a mandibular construction is known only within the insects and corresponds to a monophyletic lineage comprising the silverfish (*Zygentoma*) and all winged orders (that is, Pterygota)^{6,7}. Thus, *Rhyniognatha* is definitively an insect. The stoutly sclerotized mandibles are surrounded by a suite of other sclerotized debris. Various tissues appear to represent fragments of the head capsule, whereas others ventral to and extending beyond the mandibular apices appear to represent maxillary fragments (perhaps portions of laciniae; Fig. 1). Extending from the base of the mandibles are large fan-like structures, which may have been portions of apodemes or

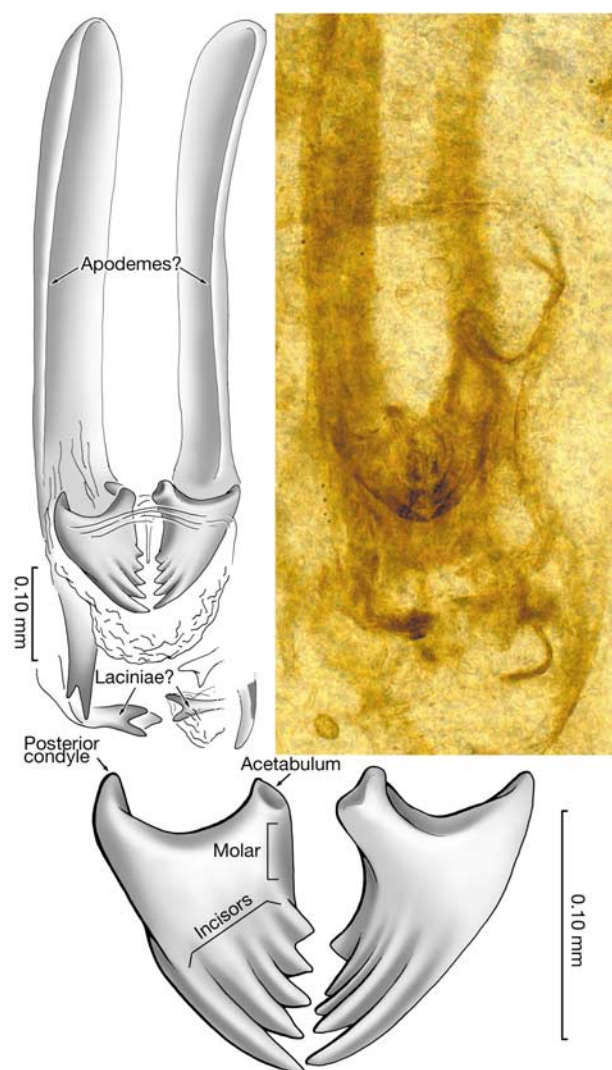


Figure 1 Remains of the oldest fossil insect. Holotype of *Rhyniognatha hirsti*, preserved in Early Devonian (Pragian) chert of Rhynie, Scotland. Photomicrograph of holotype and rendering of preserved remains, with enlarged detail of mandibles.

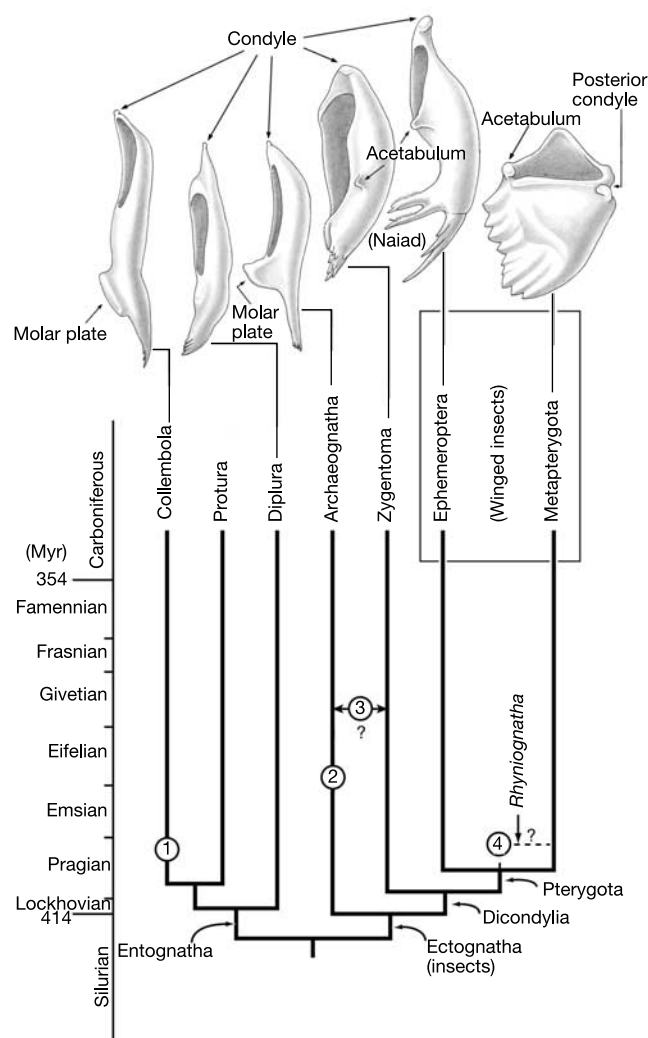


Figure 2 Phylogeny of basal hexapod orders¹⁶. Mandibular structure across the orders and putative position of *R. hirsti* (4) and other Devonian hexapods indicated (1, *Rhyniella praecursor* from Rhynie, Scotland; 2, putative bristletail from Gaspé, Quebec; 3, bristletail or silverfish remains from Gilboa, New York). Metapterygota (the clade consisting of Odonoptera and Neoptera) is represented by an orthopteran mandible.

perhaps even components of a tentorium. A more elaborate treatment of the surrounding sclerites will be presented elsewhere. Most remarkable is that the mandibles are short and triangular, rather than the elongate mandibles of Entognatha, Archaeognatha and Zygentoma (Fig. 2). In fact, this short and triangular morphology is typical of some Pterygota, suggesting that *Rhyniognatha* may have been a winged insect, although the definitive apomorphy, wings, are missing.

The most plesiomorphic lineage of winged insects are the mayflies (Ephemeroptera), which retain a subimaginal moult, very long cerci, and a median caudal filament, among other traits. Adult mayflies have vestigial mouthparts; however, naiads have generally primitive mandibles: these are elongate, with the posterior condyle (homologous with the articulation in Archaeognatha) set posteriorly along an extension of the mandible. The anterior articulation of mayflies is, by contrast to other dicondylid, winged insects, relatively weakly developed and as a result functions more similarly to a hinged joint^{14,15}. The complete fixation of the anterior articulation by a ridged ball-socket joint is a defining feature of the Metapterygota^{14–16} (that is, Odonoptera and Neoptera). The mandibles of *Rhyniognatha* also surprisingly resemble metapterygotes by the close position of the posterior condyle to the anterior acetabulum, the latter structure being well developed (Fig. 2) and distinctly unlike the weakened hinge of mayflies. *Rhyniognatha* consists of dicondylid mandibles that are not elongate, but are short and triangular as in basal metapterygote orders, and differs from basal insects (Ectognatha) as follows: from Entognatha and Archaeognatha it differs by the dicondylid mandibular structure; from Zygentoma it differs by the shortened and triangular metapterygotean construction of the mandible, discontinuous molar and incisor areas, and absence of a molar brush (also absent in Lepidodromidae); it further differs from Ephemeroptera by the functional ball-and-socket joint of the anterior articulation (Fig. 1). The incisor and molar areas of *Rhyniognatha* are well differentiated, the incisor bearing stout, sharp teeth and the molar formed by a broad toothless region (Fig. 1). An anterior acetabulum (forming the anterior articulation) does not occur in any basal hexapod order except silverfish (Zygentoma) and the winged insects, hence Hennig's recognition of the latter two as a group called Dicondylia⁶. The diversity of insect lifestyles and diets is reflected in the myriad functional types of insect mouthparts¹⁷. Given the generalized structure of *Rhyniognatha*'s mandibles and presumed other mouthpart fragments this was clearly a chewing insect, but whether its diet was spores/pollen, leaf/stem tissue or small animals is impossible to say.

The metapterygotean structure of *Rhyniognatha*'s mandibles has profound implications regarding insect evolution. Current fossil evidence indicates that insect wings originated in the Early Carboniferous period, some 90, 170 and 270 Myr before pterosaurs, birds and bats, respectively, and these structures are believed to have led largely to the spectacular diversification of insects. Fossil insects are completely absent from the Late Devonian and Early Carboniferous, and a significant diversity of palaeopterous and neopterous species appeared suddenly in the earliest Late Carboniferous^{18,19}. Besides the derived structure of *Rhyniognatha*'s mandibles, further support for winged insects 80 Myr earlier than previously known pterygotes is provided by a recent DNA study²⁰. That study estimated that insects originated 434 Myr ago (in the Early Silurian) and pterygotes originated 387 Myr ago (in the mid-Devonian) (note, however, that only neopteran species were used to represent the entire pterygote lineage, and the dating actually applies to neoptery rather than wings, which must be older).

A Devonian origin of winged insects is highly relevant to several current hypotheses on the origin of wings. One hypothesis is that metabolically expensive insect flight evolved in hyperoxic atmospheres similar to what occurred in the Carboniferous and Permian²¹, but the oxygen content of Devonian atmospheres was far

less, approximately 15%²². Also, the paranotal theory hypothesizes that insect wings evolved from lateral extensions of the thorax, called paranotal lobes (but see also the exite theory²³), originally used for controlled gliding similar to what modern silverfish are capable of²⁴. Under this hypothesis, paranotal lobes presumably evolved later into broader, hinged structures capable of powered flight, analogous to the transitional forms seen with *Archaeopteryx* and modern birds²⁵ and with 'flying lemurs' (Dermoptera) and bats²⁶. Gliding requires a perch, and there is abundant evidence that various Palaeozoic insects grazed on spores from sporangia²⁷, which in most Devonian plants were produced at the branch tips^{28,29}. In fact, all of these spore feeders had generalized, chewing mandibles, not unlike those of *Rhyniognatha*. If winged insects appeared in the Early Carboniferous this would be approximately 30 Myr after arborescent plants evolved^{28,29}, but *Rhyniognatha* occurred as tracheophytes became shrubby plants approximately 1 m tall^{28,29}. If *Rhyniognatha* had wings, for which we believe evidence is compelling, then the chronology for the evolution of wings would be more similar to that for arborescence.

The most definitive implication of *Rhyniognatha*'s dicondylid is that insects and hexapods as a whole must have originated during the Silurian period. Monocondylid insects must be older than *Rhyniognatha* (Fig. 2), and may have had an earlier start on terrestrial life than previously believed. Plants are now understood to have ventured onto land in the Ordovician period³⁰, and it is possible that amphibious invertebrates did so as well. Silurian terrestrial biotas probably included wingless insects as well as various chelicerate and myriapodous arthropods. Thus, the currently pervasive insects might not have been just the first organisms to take to the air with controlled flight (sometime during the Devonian period) but also may have been members of the earliest terrestrial faunas. □

Received 23 September; accepted 9 December 2003; doi:10.1038/nature02291.

- Rice, C. M. et al. A Devonian auriferous hot spring system, Rhynie, Scotland. *J. Geol. Soc. Lond.* **152**, 229–250 (1995).
- Hirst, S. & Maulik, S. On some arthropod remains from the Rhynie Chert (Old Red Sandstone). *Geol. Mag.* **63**, 69–71 (1926).
- Whalley, P. & Jarzembowski, E. A. A new assessment of *Rhyniella*, the earliest known insect, from the Devonian of Rhynie, Scotland. *Nature* **291**, 317 (1981).
- Labandeira, C. C., Beall, B. S. & Hueber, F. M. Early insect diversification: Evidence from a Lower Devonian bristletail from Québec. *Science* **242**, 913–916 (1988).
- Shear, W. A. et al. Early land animals in North America: Evidence from Devonian age arthropods from Gilboa, New York. *Science* **224**, 492–494 (1984).
- Hennig, W. *Insect Phylogeny* (John Wiley & Sons, New York, 1981).
- Kristensen, N. P. in *The Insects of Australia* (ed. Naumann, I. D.) 125–140 (Cornell Univ. Press, Ithaca, New York, 1991).
- Jeram, A. J., Selden, P. A. & Edwards, D. Land animals in the Silurian: Arachnids and myriapods from Shropshire, England. *Science* **250**, 658–661 (1990).
- Tillyard, R. J. Some remarks on the Devonian fossil insects from the Rhynie chert beds, Old Red Sandstone. *Trans. Entomol. Soc. Lond.* **76**, 65–71 (1928).
- Scourfield, D. J. The oldest known fossil insect. *Nature* **145**, 799–801 (1940).
- Carpenter, F. M. in *Treatise on Invertebrate Paleontology, Part R: Arthropoda 4: Superclass Hexapoda* (ed. Kaesler, R. L.) 1–655 (Geol. Soc. Am., Boulder, Colorado, 1992).
- Rasnitsyn, A. P. & Quicke, D. L. J. *History of Insects* (Kluwer, Dordrecht, Netherlands, 2002).
- Kühne, W. G. & Schlüter, T. A fair deal for the Devonian Arthropoda fauna of Rhynie. *Entomol. Gen.* **11**, 91–96 (1985).
- Staniczek, A. H. The mandible of silverfish (Insecta: Zygentoma) and mayflies (Ephemeroptera): Its morphology and phylogenetic significance. *Zool. Anz.* **239**, 147–178 (2000).
- Fürst von Lieven, A. The transformation from monocondylous to dicondylous mandibles in the Insecta. *Zool. Anz.* **239**, 139–146 (2000).
- Wheeler, W. C., Whiting, M., Wheeler, Q. D. & Carpenter, J. M. The phylogeny of the extant hexapod orders. *Cladistics* **17**, 113–169 (2001).
- Labandeira, C. Insect mouthparts: Ascertaining the paleobiology of insect feeding strategies. *Annu. Rev. Ecol. Syst.* **28**, 153–193 (1997).
- Brauckmann, C., Brauckmann, B. & Gröning, E. The stratigraphical position of the oldest known Pterygota (Insecta, Carboniferous, Namurian). *Ann. Soc. Geol. Belg.* **117**, 47–56 (1996).
- Grimaldi, D. Insect evolutionary history from Handlirsch to Hennig, and beyond. *J. Paleontol.* **75**, 1152–1160 (2001).
- Gaunt, M. W. & Miles, M. A. An insect molecular clock dates the origin of the insects and accords with palaeontological and biogeographic landmarks. *Mol. Biol. Evol.* **19**, 748–761 (2002).
- Dudley, R. *The Biomechanics of Insect Flight: Form, Function, Evolution* (Princeton Univ. Press, Princeton, New Jersey, 2000).
- Berner, R. A. & Canfield, D. E. A new model for atmospheric oxygen over Phanerozoic time. *Am. J. Sci.* **289**, 333–361 (1989).
- Kukalová-Peck, J. in *The Insects of Australia* (ed. Naumann, I. D.) 141–179 (Cornell Univ. Press, Ithaca, New York, 1991).

24. Hasenfuss, I. A possible evolutionary pathway to insect flight starting from lepidopteran organization. *J. Zool. Syst. Evol. Res.* **40**, 65–81 (2002).
25. Chiappe, L. The first 85 million years of avian evolution. *Nature* **378**, 349–355 (1995).
26. Novacek, M. J. Fossils, topologies, missing data, and the higher level phylogeny of eutherian mammals. *Syst. Biol.* **41**, 58–73 (1992).
27. Edwards, D., Selden, P. A., Richardson, J. B. & Axe, L. Coprolites as evidence for plant–animal interactions in Siluro-Devonian terrestrial ecosystems. *Nature* **377**, 329–331 (1995).
28. Gensel, P. G. & Andrews, H. N. *Plant Life in the Devonian* (Praeger, New York, 1984).
29. Alego, T. J., Scheckler, S. E. & Maynard, J. B. in *Plants Invade the Land* (eds Gensel, P. G. & Edwards, D.) 213–236 (Columbia Univ. Press, New York, 2001).
30. Wellman, C. H., Osterloff, P. L. & Mohiuddin, U. Fragments of the earliest land plants. *Nature* **425**, 282–285 (2003).

Acknowledgements We thank A. J. Ross for loan of the material discussed herein and M. G. Rightmyer for comments on and support of the study.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.S.E. (msengel@ku.edu).

Diclofenac residues as the cause of vulture population decline in Pakistan

J. Lindsay Oaks¹, Martin Gilbert², Munir Z. Virani², Richard T. Watson², Carol U. Meteyer³, Bruce A. Rideout⁴, H. L. Shivaprasad⁵, Shakeel Ahmed⁶, Muhammad Jamshed Iqbal Chaudhry⁶, Muhammad Arshad⁶, Shahid Mahmood⁶, Ahmad Ali⁶ & Aleem Ahmed Khan⁶

¹Department of Veterinary Microbiology and Pathology, Washington State University, Pullman, Washington 99164-7040, USA

²The Peregrine Fund, 5668 West Flying Hawk Lane, Boise, Idaho 83709, USA

³USGS–National Wildlife Health Center, 6006 Schroeder Road, Madison, Wisconsin 53711-6223, USA

⁴Center for Reproduction of Endangered Species, Zoological Society of San Diego, PO Box 120551, San Diego, California 92112, USA

⁵California Animal Health and Food Safety Laboratory System–Fresno Branch, University of California at Davis, 2789 S. Orange Avenue, Fresno, California 93725, USA

⁶Zoology Division, Institute of Pure and Applied Biology, Bahauddin Zakariya University, Multan, Pakistan

The Oriental white-backed vulture (OWBV; *Gyps bengalensis*) was once one of the most common raptors in the Indian subcontinent¹. A population decline of >95%, starting in the 1990s, was first noted at Keoladeo National Park, India². Since then, catastrophic declines, also involving *Gyps indicus* and *Gyps tenuirostris*, have continued to be reported across the subcontinent³. Consequently these vultures are now listed as critically endangered by BirdLife International⁴. In 2000, the Peregrine Fund initiated its Asian Vulture Crisis Project with the Ornithological Society of Pakistan, establishing study sites at 16 OWBV colonies in the Kasur, Khanewal and Muzaffargarh–Layyah Districts of Pakistan to measure mortality at over 2,400 active nest sites⁵. Between 2000 and 2003, high annual adult and subadult mortality (5–86%) and resulting population declines (34–95%) (ref. 5 and M.G., manuscript in preparation) were associated with renal failure and visceral gout. Here, we provide results that directly correlate residues of the anti-inflammatory drug diclofenac with renal failure. Diclofenac residues and renal disease were reproduced experimentally in OWBVs by direct oral exposure and through feeding vultures diclofenac-treated livestock. We propose that residues of veterinary diclofenac are responsible for the OWBV decline.

Between 2000 and 2002 we performed gross post-mortem examinations on 259 adult and subadult OWBVs, of which 219 (85%) had grossly apparent urate deposits, characteristic of visceral gout, on the surface of internal organs (Fig. 1). Visceral gout in birds is most commonly the result of renal failure leading to hyperuricaemia and the deposition of uric acid on and within the internal organs, and can be caused by degenerative, metabolic, infectious, or toxic diseases⁶. To verify renal disease, and to determine the cause, detailed necropsies and diagnostic testing were performed on a subset of 42 OWBVs (33 adults and 9 juveniles; 28 with visceral gout and 14 without visceral gout) that were found within approximately 24 h of death. The remaining OWBVs were significantly decomposed and thus were unsuitable for diagnostic evaluation, although the grossly apparent and characteristic lesions of visceral gout allowed the presence or absence of the disease to be determined. Of the 14 OWBVs without visceral gout, we determined the cause of death in 8 (57%), which included trauma, intestinal foreign bodies, lead poisoning, organophosphate poisoning and gun-shot (Supplementary Information). We were unable to determine the cause of death in the remaining six vultures. Among the 28 OWBVs with visceral gout, only one (4%) had an identifiable disease (infection with *Mycobacterium avium*) in addition to visceral gout. All but two of the visceral gout cases were in good physical condition based on the subjective assessment of normal pectoral muscle mass and adequate body fat. In all of the OWBVs with visceral gout, the only significant histopathological lesion was severe, acute renal

Table 1 Diclofenac residue testing results in kidney samples from wild OWBVs with and without renal failure

Case no.	Date	Site	Age	Gout	Diagnosis	Diclofenac residues ($\mu\text{g g}^{-1}$)
33	2001	KS	Juv	Yes	None	0.051
74	2002	KH	Ad	Yes	None	0.054
16	2001	KS	Ad	Yes	None	0.060
53	2002	KH	Ad	Yes	None	0.064
60	2002	ML	Ad	Yes	None	0.064 (0.076)
39	2001	KH	Ad	Yes	None	0.077
69	2002	ML	Ad	Yes	None	0.079
57	2002	KH	Ad	Yes	None	0.080
40	2002	ML	Ad	Yes	None	0.091
20	2001	KS	Ad	Yes	None	0.097
15	2001	KS	Ad	Yes	None	0.099
35	2001	KS	Ad	Yes	None	0.106 (0.077)
41	2002	KH	Ad	Yes	None	0.109
75	2002	KH	Ad	Yes	None	0.114
55	2002	ML	Ad	Yes	None	0.124
54	2002	ML	Ad	Yes	None	0.177
71	2002	KH	Ad	Yes	None	0.179
61	2002	ML	Ad	Yes	None	0.186
42	2002	KH	Ad	Yes	None	0.199
44	2002	KH	Ad	Yes	None	0.233
4	2000	KS	Ad	Yes	<i>M. avium</i> infection	0.450
38	2001	ML	Ad	Yes	None	0.451
59	2002	ML	Ad	Yes	None	0.504
45	2002	ML	Ad	Yes	None	0.642 (0.197)
56	2002	KH	Ad	Yes	None	0.643
2	2000	LH	Ad	No	Wire collision	BDL
3	2000	CW	Juv	No	Hit by car	BDL
12	2001	ML	Ad	No	None	BDL
14	2001	KS	Ad	No	Lead poisoning	BDL
28	2001	KS	Ad	No	Normal (trapped)	BDL (BDL)
31	2001	KH	Juv	No	Fell from nest	BDL
46	2002	KH	Juv	No	Fractured tibia	BDL
47	2002	KH	Juv	No	Intestinal foreign body	BDL
49	2002	KF	Ad	No	None	BDL
50	2002	ML	Juv	No	None	BDL
51	2002	KH	Juv	No	None	BDL
52	2002	ML	Juv	No	None	BDL
58	2002	ML	Ad	No	Organophosphate	BDL (BDL)

BDL indicates 'below detection limit' of diclofenac assay ($0.005\text{--}0.01\text{ }\mu\text{g g}^{-1}$). Results in parentheses are from the Toxicology Laboratory at the University of Pennsylvania New Bolton Center, which were performed as independent verification. This additional testing also did not detect acetaminophen (0.05), flunixin (0.05), ibuprofen (0.50), phenylbutazone (0.10), oxyphenylbutazone (0.05), indomethacin (0.05), ketoprofen (0.25), mefenamic acid (0.50), salicylic acid (1.0), tolmetin (0.05), or naproxen (1.0) (detection limits, in parentheses, are in $\mu\text{g g}^{-1}$). Ad, adult; CW, Chichawatni; Juv, juvenile; KF, Katora Forest; KH, Khanewal district study site; KS, Kasur district study site; LH, City of Lahore; ML, Muzaffargarh–Layyah district study site.

tubular necrosis and uric acid crystal formation in the kidneys and other tissues (detailed histopathological description is provided in Supplementary Information). Fibrosis or other changes indicating chronic renal disease were not present. Inflammatory lesions indicating an infectious disease were not consistently present in any of the visceral gout cases.

These findings were most compatible with acute renal failure due to a toxic cause. Additional testing eliminated known causes of avian renal disease, including toxic concentrations of cadmium, lead (39 vultures) and mercury (37 vultures)⁷ (Supplementary Tables 1 and 2), as well as infection with the avian influenza⁸, infectious bronchitis⁹ (21 vultures) and West Nile viruses¹⁰ (13 vultures) (Supplementary Table 3). Testing was also negative for toxic concentrations of arsenic, and for toxic or deficient concentrations of copper, iron, manganese, molybdenum and zinc (39 vultures) (Supplementary Tables 1 and 2). Carbamate and organophosphate pesticides (34 vultures), organochlorine pesticides and polychlorinated biphenyls (13 vultures) were either not detected or were detected at below toxic concentrations (Supplementary Tables 4, 5 and 6). We failed to isolate any viruses from the kidney, spleen, lung or intestine using at least four passages on chicken embryo fibroblast, duck embryo fibroblast, peregrine falcon (*Falco peregrinus*) embryo fibroblast and Vero cell cultures for 13 vultures (Supplementary Table 7).

The primary food source for OWBVs in Pakistan is dead domestic livestock. Therefore, we hypothesized that ingested veterinary pharmaceuticals might be responsible for renal disease in the scavenging birds. We conducted a survey of 74 veterinarians and veterinary pharmaceutical retailers in the region to identify drugs that were known to be both toxic to kidneys and absorbed orally. The only drug identified in the survey that met these criteria was diclofenac, a non-steroidal anti-inflammatory drug (NSAID) used as an analgesic, anti-inflammatory and antipyretic¹¹. Liquid chromatography and mass spectroscopy detected diclofenac residue concentrations of 0.051–0.643 $\mu\text{g g}^{-1}$ in the kidneys of 25 out of 25 (100%) vultures that died of renal failure, and in 0 out of 13 (0%) vultures that died from other causes (Table 1). The association between renal failure and diclofenac residues was very highly significant ($\chi^2 = 38$, 1 degree of freedom, $P = 0.0$). Independent testing on three positive and two negative samples verified these results, and did not detect the presence of any other NSAID

(Table 1). Diclofenac residues were found in vultures collected between 2000 and 2002, and at study sites separated by up to 350 km.

Although there are no reports of the use of diclofenac in birds, other NSAIDs such as indomethacin¹², flunixin meglumide¹³ and ketoprofen (M. Busch & B.A.R., personal communication) may cause renal disease in chickens, cranes and quail, and in African white-backed vultures (*Gyps africanus*), respectively. NSAIDs, including diclofenac, are also recognized as causing renal disease in mammals^{11,14}. To verify the renal toxicity of diclofenac in OWBVs, we administered single oral doses of veterinary diclofenac to four captive, non-releasable juvenile OWBVs. Two were given the mammalian label dose of 2.5 mg kg^{-1} (high dose) and two were given 0.25 mg kg^{-1} (low dose). Both of the high-dose and one of the low-dose vultures died as a result of renal failure and visceral gout within 36–58 h after administration, and all had the same microscopic renal lesions as the field cases. Plasma samples collected 1 h after administration for one high-dose and one low-dose vulture indicated normal uric acid levels¹⁵ (53 and 33 mg l^{-1} , respectively). By 24 h after administration, both vultures had developed hyperuricaemia (775 mg l^{-1} for both). Diclofenac residue concentrations of 0.29, 1.1 and 0.16 $\mu\text{g g}^{-1}$ were present in the kidneys of the two high-dose vultures and the one low-dose vulture that died as a result of renal failure. Although the other low-dose vulture developed hyperuricaemia (138 mg l^{-1} at 1 h and 654 mg l^{-1} at 24 h after administration), it remained clinically normal 4 weeks after administration, and did not have microscopic renal lesions or detectable diclofenac residues at necropsy. Two untreated control birds that were fed the same diet also did not have renal lesions or detectable diclofenac residues at necropsy.

The most probable source of diclofenac exposure is the consumption of treated livestock. In Pakistan, veterinary diclofenac is widely marketed over the counter by multiple companies for treating all types of hoofed livestock. A second survey of 84 drug retailers and veterinarians in nine districts in the Punjab province (Dera Ghazi Khan, Kasur, Khanewal, Layyah, Multan, Muzaffargarh, Pakpattan, Rajanpur, Sahiwal) and one district in the Sindh province (Hyderabad) found that all 84 sold diclofenac; 77 sold it daily, and 71 reported that it had become available within the last 5 years. Furthermore, livestock that die of disease or injury are typically left for scavengers to remove.



Figure 1 Necropsy photograph of the abdominal cavity of an OWBV with visceral gout, showing uric acid precipitates on the serosal surfaces of the liver.

To verify that carcasses of treated livestock contain sufficient diclofenac concentrations to cause renal failure and death in the vultures that scavenge upon them, ten juvenile OWBVs were fed meat from a buffalo or goat that was injected intramuscularly with 2.5 mg kg⁻¹ veterinary diclofenac once daily for 3 days (per label instructions) and that were slaughtered 4 h hours after the last injection. Resulting diclofenac residue concentrations in the buffalo kidney, liver and muscle were 5.7, 1.5 and 0.76 µg g⁻¹, respectively, and in the goat kidney, liver and muscle they were 0.94, 0.22 and 0.19 µg g⁻¹, respectively. Ten more OWBVs were fed buffalo meat that contained 6.4 µg g⁻¹ of diclofenac. On the basis of the product of diclofenac concentration in the meat and the amount of food consumed, eight OWBVs received doses of 0.005–0.3 mg kg⁻¹. Two of these vultures died from renal failure at 4 and 6 days after exposure, and at necropsy they had diclofenac residue concentrations of 0.07 and 0.38 µg g⁻¹ in the kidney. A necropsy was carried out on one surviving, clinically normal vulture at 8 days after exposure, at which time it did not have renal lesions or detectable diclofenac residues. The other five surviving vultures remain clinically normal at approximately 6 months after exposure. Two OWBVs received doses of 0.5–0.6 mg kg⁻¹. One of these died from renal failure 1 day after exposure, and at necropsy it had a diclofenac residue concentration of 0.25 µg g⁻¹ in the kidney. The other surviving, clinically normal vulture had a necropsy performed on it at 8 days after exposure, at which time it did not have renal lesions or detectable diclofenac residues. Ten OWBVs received doses of 0.8–1.0 mg kg⁻¹. All ten of these vultures died from renal failure. Kidney diclofenac residue concentrations of 0.120–0.906 µg g⁻¹ were present in six out of six of these vultures tested. These ten vultures were negative for other recognized avian nephrotoxins (Supplementary Information). All 13 vultures that died from renal failure after consuming meat from diclofenac-treated animals had the same histopathological renal lesions as the vultures that were exposed directly with diclofenac, and as the field cases with visceral gout. Six control vultures that were fed untreated goat (meat that did not contain detectable diclofenac residues) remained clinically normal 8 and 25 days into the experiment. Two were necropsied at day eight, and they did not have renal lesions or detectable

diclofenac residues. These results are summarized in Table 2. In addition, all of the experimental OWBVs were held in captivity, and were clinically normal for 30–75 days before diclofenac exposure, indicating that visceral gout was not incidentally related to captivity. The combined mortality rate of 13 out of 20 (65%) in the exposed vultures and 0 out of 6 (0%) in the control vultures indicates a statistically highly significant relationship between renal failure and exposure to diclofenac ($\chi^2 = 7.8$, 1 degree of freedom, $P = 0.005$). When the direct and food-borne exposure levels were combined, a statistically highly significant ($\chi^2 = 12.3$, 3 degrees of freedom, $P = 0.0064$) dose-dependence for diclofenac toxicity can also be demonstrated.

Alternatively, wild OWBVs might be exposed to diclofenac through contaminated water sources¹⁶; however, because NSAID-associated renal failure in birds appears to be acute and dose-dependent^{12,13}, the very low water concentrations (on the scale of ng l⁻¹)¹⁶ would be unlikely to cause toxicity. Also, the rapid and complete elimination of most NSAIDs in mammals¹⁴ and birds¹⁷ makes bioaccumulation in either the treated animals or in the scavengers that feed on them unlikely. The lack of detectable diclofenac residues in the kidneys of any unaffected wild OWBV and in the kidneys of the surviving experimental vultures also suggest that bioaccumulation does not occur.

The potential ecologically toxic effects of human and veterinary pharmaceuticals are a growing concern^{16,18}. Sporadic poisonings of scavenging birds by organophosphate pesticides and barbiturates used in livestock have been documented previously^{19,20}. Anthelmintics may be shed in livestock faeces and may kill ecologically important invertebrates²¹. However, in stark contrast to the current situation with the OWBV, pharmaceutical residues have not been implicated previously as a cause of major ecological damage. The identification of diclofenac as the cause of the OWBV decline in Pakistan provides an opportunity for conservation intervention. The high rate of visceral-gout-associated vulture mortality in India^{3,22,23} as well as the widespread use of veterinary diclofenac in India (R. Risebrough, personal communication) suggests strongly that diclofenac may also be responsible for vulture declines in the rest of the Indian subcontinent wherever diclofenac is used for the treatment of livestock. □

Methods

Histopathology

Tissue samples were fixed in 10% neutral-buffered formalin, and 5-µm sections were cut and stained routinely with haematoxylin and eosin.

Toxicology

Testing for arsenic, cadmium, copper, iron, lead, manganese, mercury, molybdenum and zinc was performed by the inductively coupled plasma method. Organophosphate and carbamate testing included measuring brain acetylcholinesterase levels²⁴ as an indirect indicator of exposure²⁵, and direct detection by liquid chromatography and mass spectroscopy on liver, brain or stomach contents. Organochlorine pesticide and polychlorinated biphenyl testing was performed by liquid chromatography and mass spectroscopy. Reported concentrations are based on the wet weight of the samples.

Diclofenac analysis

Kidney samples (0.5 g) were homogenized in 4 ml of acetonitrile and centrifuged to pellet debris. Three millilitres of the supernatants were then passed through solid-phase extraction cartridges (Waters Sep-Pak Plus t-C18) at 1 ml min⁻¹. The cartridges were then washed with an additional 3 ml of acetonitrile, the eluates pooled, and concentrated to final volumes of 1 ml. For samples that weighed less than 0.5 g, the protocol was adjusted proportionally for the smaller weights.

Diclofenac was detected and quantified by high-performance liquid chromatography and mass spectroscopy (Agilent 1100 series equipped with a Waters Xterra MS C18 (3.9 mm × 150 mm, 5 µm) and guard column (3.9 mm × 20 mm, 5 µm)). Diclofenac standard (Sigma, D6899) was dissolved into 1:1 (v/v) acetonitrile:water at 1,040 µg ml⁻¹, diluted to a final concentration of 10.4 µg ml⁻¹, and this stock was used to prepare calibration standards of 0.104, 0.052, 0.026, 0.0104 and 0.0052 µg ml⁻¹ in acetonitrile, which generated a linear calibration curve with R^2 values equal to 0.999. Samples and standards (20 µl) were subjected to a binary gradient elution profile, which consisted of 0.1% acetic acid in water (solution A) and 100% acetonitrile (solution B) as follows: starting conditions 75% A/25% B for 0.1 min, a 15-min linear gradient from 75% A/25% B to 5% A/95% B, followed by a 5-min column-wash step in 5% A/95% B, and a 10-min

Table 2 Results of experimentally feeding meat from diclofenac-treated animals to OWBVs

Vulture no	Exposure (mg kg ⁻¹)	Result	Diclofenac residues (µg g ⁻¹)
11	0.007	Gout	0.38
12	0.025	Survived	ND
8	0.027	Survived	ND
10	0.028	Survived	ND
7	0.029	Survived	ND
9	0.029	Survived	ND
D	0.140	Gout	0.07
F	0.240	Survived	BDL (8 days)
H	0.550	Gout	0.25
B	0.600	Survived	BDL (8 days)
64	0.820	Gout	0.54
65	0.820	Gout	ND
62	0.840	Gout	0.22
63	0.840	Gout	0.12
70	0.860	Gout	0.74
67	0.860	Gout	ND
68	0.860	Gout	ND
66	0.880	Gout	0.26
73	0.910	Gout	0.91
72	0.940	Gout	ND
1	None (control)	Survived	ND
2	None (control)	Survived	ND
3	None (control)	Survived	ND
4	None (control)	Survived	ND
cc1	None (control)	Survived	BDL (8 days)
cc2	None (control)	Survived	BDL (8 days)

Values in parentheses in the last column indicate time after exposure. OWBVs that died and displayed gross characteristics of visceral gout on necropsy are labelled as 'Gout' in the third column. BDL, below detection limit of diclofenac assay (0.005–0.01 µg g⁻¹); ND, analysis not done.

re-equilibration step with 75% A/25% B before the next injection. Flow rate was 0.7 ml min⁻¹ and the column temperature was 40 °C. The mass spectrometer (Agilent 1946D) was equipped with an electrospray ionization inlet and mass spectra were acquired in the negative ion mode. The mass spectrometer was set to selectively monitor for mass ions 294 and 296 *m/z* of which mass 294 *m/z*, the deprotonated molecular mass, was used to quantify the diclofenac residues. The mean recovery from spiked avian kidney tissue was approximately 70%, and this value was used to calculate the final sample concentration of diclofenac.

Virus isolation

Virus isolation in cell culture was performed in primary avian cell cultures prepared by the method of ref. 26 from chicken, duck and peregrine falcon embryos, and in Vero cells. Tissue samples were prepared by homogenization in Eagle's minimal essential medium with penicillin, streptomycin and amphotericin B, centrifuged to remove debris, filtered through 0.45-µm filters, and added to the cells. Cell cultures were observed daily for cytopathology. At the end of each passage (5–8 days), the cells and media were aspirated, stored at –80 °C, rapidly thawed, and added to new cells for the next passage.

Polymerase chain reaction assays for infectious agents

RNA and DNA were extracted from 0.1–0.3 g of tissue with a commercial RNA extraction kit (Trizol Reagent; BRL Life Technologies) or a commercial DNA extraction kit (Puregene; Gentra Systems) as per the manufacturer's instructions. In addition to positive and negative control RNA or DNA for each agent, intact sample RNA and DNA were verified by demonstrating the presence of avian cellular β-actin gene RNA or DNA, and the absence of nonspecific inhibitors was verified by spiking samples with positive control RNA or DNA.

The polymerase chain reaction with the reverse transcription (RT–PCR) procedure for infectious bronchitis virus was as published in ref. 27—the positive control for this assay was the Arkansas strain. The RT–PCR procedure for avian influenza was as published in ref. 28—the positive control for this assay was viral RNA obtained from the USDA National Veterinary Services Laboratory, Ames, Iowa. The RT–PCR procedure for West Nile virus was as published in ref. 29—the positive control for this assay was viral RNA obtained from the USDA National Veterinary Services Laboratory, Ames, Iowa.

Identification of *M. avium*

Acid-fast bacilli were detected with Ziehl–Neelsen stains of tissue impression smears. *Mycobacterium avium* was identified by PCR amplification and restriction enzyme polymorphisms in the heat-shock protein gene³⁰.

Uric acid measurement

Uric acid levels were measured in heparinized plasma with a colorimetric assay for the production of hydrogen peroxide by uricase.

Animal subjects

Animals were maintained and experiments performed in accordance with the Animal Health and Welfare Regulations of Bahauddin Zakariya University, Pakistan.

Received 14 September; accepted 23 December 2003; doi:10.1038/nature02317.

Published online 28 January 2004.

1. Ali, S. & Ripley, S. D. *Handbook of the Birds of India and Pakistan Together With Those of Nepal, Bhutan, and Ceylon I. Divers to Hawks* (Oxford Univ. Press, Oxford, 1968).
2. Prakash, V. Status of vultures in Keoladeo National Park, Bharatpur, Rajasthan, with special reference to population crash in *Gyps* species. *J. Bombay Nat. Hist. Soc.* **96**, 365–378 (1999).
3. Pain, D. *et al.* Causes and effects of temporospatial declines of *Gyps* vultures in Asia. *Conserv. Biol.* **17**, 661–671 (2003).
4. BirdLife International. *Threatened Birds of Asia: the BirdLife International Red Data Book* (BirdLife International, Cambridge, 2001).
5. Gilbert, M. *et al.* Breeding and mortality of Oriental White-backed vulture *Gyps bengalensis* in Punjab Province, Pakistan. *Bird Conserv. Int.* **12**, 311–326 (2002).
6. Crespo, R. & Shivaprasad, H. L. in *Diseases of Poultry* 11th edn (ed. Saif, Y. M.) 1085–1087 (Iowa State Press, Ames, 2003).
7. Beyer, W., Heinz, G. & Redmon-Norwood, A. (eds) *Environmental Contaminants in Wildlife. Interpreting Tissue Concentrations* Ch. 10, 11, 14, 17 (CRC Press, Boca Raton, 1996).
8. Swayne, D. E. & Slemons, R. D. Comparative pathology of a chicken-origin and two-duck origin influenza virus isolates in chickens: the effect of route of inoculation. *Vet. Pathol.* **31**, 237–245 (1994).
9. Ziegler, A. F. *et al.* Nephropathogenic infectious bronchitis in Pennsylvania chickens 1997–2000. *Avian Dis.* **46**, 847–858 (2002).
10. Steele, K. E. *et al.* Pathology of fatal West Nile virus infections in native and exotic birds during the 1999 outbreak in New York City, New York. *Vet. Pathol.* **37**, 208–224 (2000).
11. Todd, P. A. & Sorkin, E. M. Diclofenac sodium. A reappraisal of its pharmacodynamic and pharmacokinetic properties, and therapeutic efficacy. *Drugs* **35**, 244–285 (1988).
12. Nys, Y. & Rzaia, J. Increase in uricemia induced by indomethacin in hens or chickens. *C. R. Séances Acad. Sci. III* **296**, 401–404 (1983).
13. Paul-Murphy, J. & Ludders, J. W. Avian analgesia. *Vet. Clin. N. Am. Exotic Anim. Prac.* **4**, 35–45 (2001).
14. Murray, M. D. & Brater, D. C. Renal toxicity of the nonsteroidal anti-inflammatory drugs. *Annu. Rev. Pharmacol. Toxicol.* **33**, 435–465 (1993).
15. Gatome, C. W. *Haematology and Blood Biochemistry in Free-Living African White-Backed Vultures Gyps africanus in Kenya* (MSc, Univ. London, 2002).
16. Tixier, C., Singer, H. P., Oellers, S. & Muller, S. R. Occurrence and fate of carbamazepine, clobicric acid, diclofenac, ibuprofen, ketoprofen, and naproxen in surface waters. *Environ. Sci. Technol.* **37**, 1061–1068 (2003).
17. Baert, K. & De Backer, P. Comparative pharmacokinetics of three non-steroidal anti-inflammatory drugs in five bird species. *Comp. Biochem. Physiol. C* **134**, 25–33 (2003).

18. Daughton, C. G. & Ternes, T. A. Pharmaceuticals and personal care products in the environment: agents of subtle change. *Environ. Health Perspect.* **107**, 907–938 (1999).
19. Henny, C. J., Blas, L. J., Kolbe, E. J. & Fitzner, R. E. Organophosphate insecticide (famphur) topically applied to cattle kills magpies and hawks. *J. Wildl. Mgmt* **49**, 648–658 (1985).
20. Langelier, K. M. in *Raptor Biomedicine* (eds Redig, P. T., Cooper, J. E., Rempke, J. D. & Hunter, D. B.) 231–232 (Univ. Minnesota, Minneapolis, 1993).
21. McKellar, Q. A. Ecotoxicology and residues of anthelmintic compounds. *Vet. Parasitol.* **72**, 413–435 (1997).
22. Cunningham, A. A. *et al.* Indian vultures: victims of an infectious disease epidemic? *Anim. Conserv.* **6**, 189–197 (2003).
23. Mishra, S. K. *et al.* Vulture mortality: pathological and microbiological investigations. *Indian J. Anim. Sci.* **72**, 283–286 (2002).
24. Ellman, G. L., Courtney, K. D., Andres, V. & Featherstone, R. M. A new and rapid colorimetric determination of acetylcholinesterase activity. *Biochem. Pharmacol.* **7**, 88–95 (1961).
25. Hill, E. F. & Fleming, W. J. Anticholinesterase poisoning of birds: field monitoring and diagnosis of acute poisoning. *Environ. Toxicol. Chem.* **1**, 27–38 (1982).
26. Docherty, D. E. & Slota, P. G. Use of muscovy duck embryo fibroblasts for the isolation of viruses from wild birds. *J. Tissue Cult. Methods* **11**, 165–170 (1988).
27. Kwon, H. M., Jackwood, M. W. & Gelb, J. Differentiation of infectious bronchitis virus serotypes using polymerase chain reaction and restriction fragment length polymorphism analysis. *Avian Dis.* **37**, 194–202 (1993).
28. Lee, M. S., Chang, P., Shien, J., Cheng, M. & Shieh, H. K. Identification and subtyping of avian influenza viruses by reverse transcription-PCR. *J. Virol. Methods* **97**, 13–22 (2001).
29. Johnson, D. J., Ostlund, E. N., Pedersen, D. D. & Schmitt, B. J. Detection of North American West Nile virus in animal tissue by a reverse transcription-nested polymerase chain reaction assay. *Emerg. Inf. Dis.* **7**, 739–741 (2001).
30. Mendenhall, M. K. *et al.* Detection and differentiation of *Mycobacterium avium* and *Mycobacterium genavense* by polymerase chain reaction and restriction enzyme digestion analysis. *J. Vet. Diagn. Invest.* **12**, 57–60 (2000).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements The authors thank S. Pritchard, S. Donahoe and M. Asim for technical assistance, and T. Taruscio and HEJ Laboratories (Karachi) for performing diclofenac assays. We thank V. Beasley, K. Beckman, P. Benson, T. Besser, T. Cade, K. Mealey, R. Poppenga, P. Talcott, The Ornithological Society of Pakistan, Punjab Department of Wildlife and Parks, the National Council for the Conservation of Wildlife (Islamabad), World Wildlife Fund (Pakistan), and Brigadier M. Ahmed for their cooperation. This research was conducted as part of The Peregrine Fund's Asian Vulture Crisis Project and was supported by the Gordon and Betty Moore Foundation, The Peregrine Fund, Disney Wildlife Conservation Fund, San Diego Zoological Society, and the United Nations, Ivorybill and Summit Foundations.

Authors' contributions J.L.O. coordinated the overall diagnostic work and performed microbiology; M.G., M.Z.V. and A.A.K. coordinated field biology and collected samples; R.T.W. supervised field biology; C.U.M., B.A.R. and H.L.S. performed pathology; and S.A., M.J.I.C., M.A. and A.A. collected biology data and samples.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.L.O. (loaks@vetmed.wsu.edu).

An optimal bronchial tree may be dangerous

B. Mauroy¹, M. Filoche^{1,2}, E. R. Weibel³ & B. Sapoval^{1,2}

¹Centre de Mathématiques et de leurs Applications, Ecole Normale Supérieure de Cachan, 94235 Cachan, France

²Laboratoire de Physique de la Matière Condensée, CNRS Ecole Polytechnique, 91128 Palaiseau, France

³Department of Anatomy, University of Bern, CH-3000 Bern, Switzerland

The geometry and dimensions of branched structures such as blood vessels or airways are important factors in determining the efficiency of physiological processes. It has been shown that fractal trees can be space filling¹ and can ensure minimal dissipation^{2–4}. The bronchial tree of most mammalian lungs is a good example of an efficient distribution system with an approximate fractal structure^{5,6}. Here we present a study of the compatibility between physical optimization and physiological robustness in the design of the human bronchial tree. We show

that this physical optimization is critical in the sense that small variations in the geometry can induce very large variations in the net air flux. Maximum physical efficiency therefore cannot be a sufficient criterion for the physiological design of bronchial trees. Rather, the design of bronchial trees must be provided with a safety factor and the capacity for regulating airway calibre. Paradoxically, our results suggest that bronchial malfunction related to asthma is a necessary consequence of the optimized efficiency of the tree structure.

In the bronchial tree, airways branch by dichotomy with a systematic reduction of their length and diameter. In the human lung, for example, the conducting airway tree ends at about 2^{17} terminal bronchioles, each of which is followed by roughly six generations of alveolar ducts that constitute the acini⁷ and are involved in gas exchange. Considering the lower part of the bronchial tree (generations 6–16) and assuming that air flow in this duct system obeys Poiseuille law (a good approximation below the sixth generation at rest⁸), a ‘best’ structure can be deduced by minimizing the total viscous dissipation in a finite tree volume. A purely mathematical argument using Lagrange multipliers² suggests that the best tree is fractal with a fractal dimension equal to 3. In such an ideal tree, the successive airway segments are homothetic with a size ratio, h , equal to $(1/2)^{1/3} \approx 0.7937$. This is what is known as the Hess–Murray law, which was first formulated by Hess⁹ for blood vessels and then developed further by Murray¹⁰.

The effect of airway geometry on ventilation can be developed as follows. Assuming that bifurcating branches are symmetric (a reasonable assumption for the deeper part of the lung), the ratio of diameter and length between generation $p - 1$ and generation p is called h_p . Denoting R and V , respectively, as the resistance and volume of a given duct, the h -reduced duct has a Poiseuille resistance of R/h^3 , because this resistance is proportional to the duct length and inversely proportional to the fourth power of the diameter. By contrast, the volume is multiplied by a factor h^3 at each generation. After p generations, the sizes are reduced by a factor $h_1 \times h_2 \times \dots \times h_{p-1}$, such that the total resistance and volume of a tree with $N + 1$ generations (indexed 0 to N) can be written as:

$$R_N = R_0 + \sum_{p=1}^N \frac{1}{2^p} \frac{R_0}{(h_1 \times h_2 \times \dots \times h_p)^3}$$

$$V_N = V_0 + \sum_{p=1}^N 2^p (h_1 \times h_2 \times \dots \times h_p)^3 V_0$$

If Φ is the global airflow, the total pressure drop is $\Delta P_N = R_N \Phi$, and the total dissipation can be written as $\Phi \Delta P_N$. This power loss can be minimized relative to $h_1, \dots, h_N$ on the surface defined by the constraint $V_N = \Omega$. The minimum of R_N on $V_N = \Omega$ is characterized by the existence of a Lagrange multiplier, μ , such that

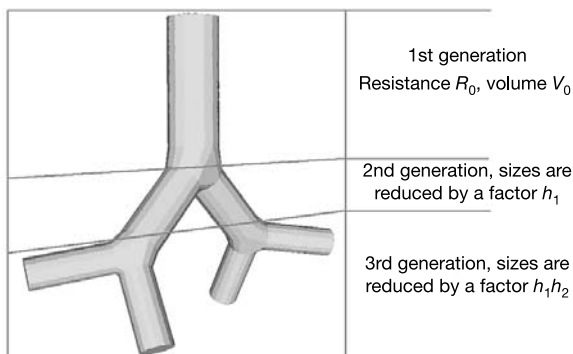


Figure 1 Geometry of a dichotomic and homothetic tree structure. In the third generation, the size reduction with respect to the first generation is the product of the homothety ratios in the second and third generations.

$\nabla(R_N) = \mu \nabla(V_N)$. The energy dissipation per unit volume of the tree, equal to $\Phi \Delta P_N / V_N$, is also optimized through this operation. This leads to $h_1 = [(\Omega - V_0)/(2NV_0)]^{1/3}$ and $h_i = (1/2)^{1/3}$ for i in the series $\{2, \dots, N\}$. The ‘best’ tree is then fractal with a constant reduction factor $h_i = (1/2)^{1/3}$. In consequence, the fractal dimension is equal to $\log(2)/\log(2^{1/3}) = 3$. For simplicity, we study here symmetric bifurcations, but, as shown below, the same results hold for asymmetric bifurcations.

More generally, assuming a constant reduction factor h between generations, the fractal dimension would be equal to $\ln(2)/\ln(1/h)$, which is larger than 3 for $h > (1/2)^{1/3}$. Figure 1 shows that with each generation the size factor falls with h , whereas the number of ducts per generation increases by a factor of 2. At generation p , therefore, the number of ducts is 2^p and the size factor is h^p . Thus, the volume and pressure drop can be written as:

$$V_N = V_0 \left[1 + \sum_{p=1}^N (2h^3)^p \right]$$

$$\Delta P_N = R_0 \Phi \left[1 + \sum_{p=1}^N \frac{1}{(2h^3)^p} \right]$$

One observes that the same factor $(2h^3)^p$ appears in both the numerator and in the denominator. Depending on whether h is smaller or larger than $(1/2)^{1/3}$, either the volume or the pressure drop is infinite for an infinitely divided tree. In other words, it is impossible to have a nonzero flux from a ‘finite’ pressure drop for an ‘infinite’ tree of finite volume. But minimizing ΔP_N for a finite tree in a given (finite) volume leads to the optimal h value equal to $(1/2)^{1/3}$.

If h is smaller than $(1/2)^{1/3} \approx 0.79$, the pressure drop for a given flux diverges with N . An increased pressure drop can be overcome by greater respiratory efforts, but in severe cases this will cause exertion. This situation can be called ‘overcritical’. In contrast, if h is larger than $(1/2)^{1/3}$, the resistance is small and the volume is larger than necessary. Notably, real lungs are not so far from optimality, with a homothety ratio h of about 0.85 (ref. 11), as shown in Fig. 2. The dependence on h of the volume and resistance of an 11-generation tree is shown in Fig. 3. Such a tree, obeying Poiseuille flow, is representative of the deeper airways in the human lung (generations 6–16). The human bronchial tree, with $h \approx 0.85$ (ref. 11), is thus not truly optimized: its volume is too large and its overall resistance is too small. This corresponds to an ‘undercritical’ condition. It gives, in fact, a safety margin for breathing with respect to possible bronchial constriction.

The fact that the homothety ratios of real bronchial trees are a few

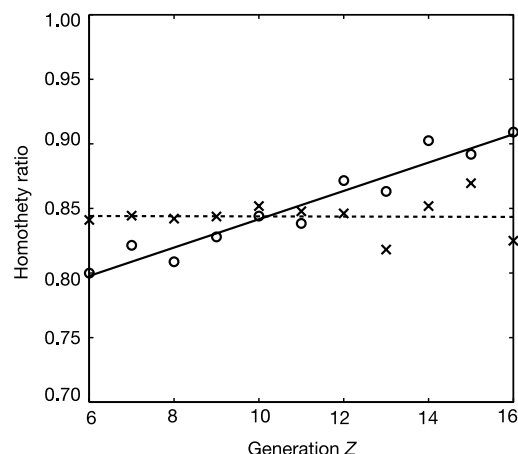


Figure 2 Homothety ratios for length and diameters of deep bronchi in the human lung ($6 \leq Z \leq 16$). Data are taken from ref. 19. Circles and crosses represent diameter and length homothety ratios, respectively.

per cent larger than the critical value has several consequences. Considered as a rigid structure, the bronchial tree would be very sensitive to geometrical imperfections. As there is always physiological variability, this would create uneven ventilation-perfusion relations in peripheral gas exchange units. This must be avoided by regulating airway calibre¹². Without such regulation, there would be a multifractal distribution of air¹³, resulting in strongly non-uniform ventilation with some regions of the lung poorly fed with fresh air. Figure 2 shows that the homothety ratio for diameter increases slightly in higher generation numbers, that is, towards the periphery. Under the hypothesis that the cast measurements reported in Fig. 2 represent the relaxed normal state of the airways, these airways allow a greater degree of contraction before a critical state is reached. In that sense, the safety margin is larger for smaller bronchioles.

In the same spirit, it should be recalled that, in asthma, bronchitis or allergenic reaction to pollen, the inner diameters of bronchioles, and not their lengths, are reduced. Thus, ducts are no longer homothetic and the diameter and length ratios of sequential bronchioles, denoted h_d and h_l , respectively, below, are altered. As shown in Fig. 2, the length reduction ratio h_l is constant and equal to 0.85, whereas h_d increases from 0.82 to 0.9. As the Poiseuille resistance is affected by the fourth power of the bronchial diameters, the pressure drop is:

$$\Delta P_N = R_0 \Phi \left[1 + \sum_{p=1}^N \frac{1}{2^p} \left(\frac{h_l}{h_d^4} \right)^p \right]$$

The fourth-power dependence of the duct resistances modifies the critical value of h_d to $(h_l/2)^{1/4} \approx 0.81$, assuming a constant value of $h_l = 0.85$.

The multiplicative effect of the tree structure combined with the fourth-power dependence of airway resistance creates an extreme sensitivity to variations in the homothety factor (see Methods). To illustrate this effect, we have computed the resistance R_{10} as a function of the h_d values when they are diminished from their normal values given in Fig. 2. The result is shown in Fig. 4. The constriction creates a marked increase in the total resistance of the tree: a 4% reduction of the homothety ratio h_d would nearly double the tree resistance, but it would increase the resistance of a simple airway tube by only 15%.

It is important to determine whether these results are applicable to the real human lung, where the bronchial tree has a marked degree of asymmetry modulating its regular features^{11,14}. Applying

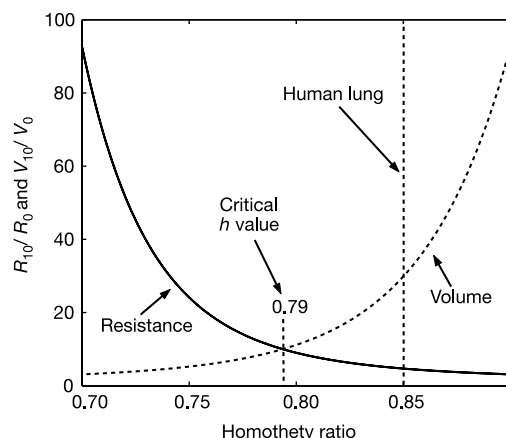


Figure 3 Dependence on the homothety ratio (h) of the resistance (R) and volume (V) of an 11-generation homothetic tree. Note the rapid variations, which indicate that the properties of the system are highly sensitive to its geometry. The homothety ratio of human airways is larger than the critical value.

the same approach, we consider an asymmetric tree built with two different reduction ratios h_1 and h_2 assigned randomly to the daughter branches at each bifurcation. Although the structure is random, the tree resistance can be written exactly as:

$$R_N = R_0 \left[1 + \sum_{p=1}^N \left(\frac{1}{h_1^3 + h_2^3} \right)^p \right]$$

The behaviour of this series is determined by the value of $1/(h_1^3 + h_2^3)$. To give an example, we have computed the effect of nonsymmetric bifurcations in which one of the branch has a systematic 20% increase in h ($h_1 = 1.2h$), whereas the other has a systematic decrease of 20% ($h_2 = 0.8h$). This corresponds to a ratio of $(1.2/0.8)^3 \approx 3.38$ between the daughter branch resistances at any bifurcation. The critical value of h is now $h = 1/(1.2^3 + 0.8^3)^{1/3} \approx 0.76$, instead of 0.79 for the symmetric model. As shown in Fig. 4, the extreme sensitivity to bronchial constriction remains (it is only slightly reduced) because the criticality depends on the basic tree structure and not on the symmetry. Notably, at criticality the asymmetric system has the same response as the symmetric tree. We conclude that other mammalian lungs that have a greater degree of asymmetry than the human lung, such as that of the dog, would behave in a similar fashion.

In summary, we have shown that the optimal system is dangerously sensitive to fluctuations or physiological variability, such that physical optimality cannot be the only criterion for the design of the bronchial tree. The morphology of the human bronchial tree is, however, close to providing maximal efficiency in assuring air distribution with minimal viscous dissipation. As the real bronchi are a little larger than optimal, they occupy a slightly larger volume than is strictly necessary. This gives the system a safety margin with respect to resistance, at the cost of an increased dead space that requires a larger tidal volume. In the human lung, the volume of the conducting airways is about 3% of the total lung volume, which is of the same order as each of the venous and arterial trees. The remaining 90% of the volume is occupied by the acini, where gas exchange takes place. The acini are themselves built by space-filling

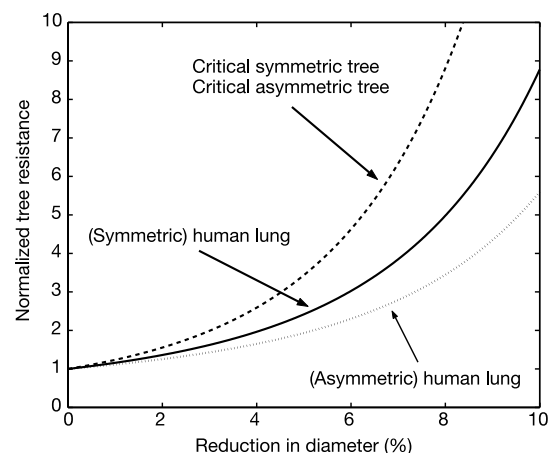


Figure 4 Dependence of tree resistance on bronchial constriction expressed as a percentage of reduction in diameter. The airway resistance has been normalized to its value without constriction. The three curves show the sensitivity of four particular trees. The dashed line corresponds to the critical trees (symmetric with $h = 0.79$ and asymmetric with $h = 0.76$). The solid line corresponds to a symmetric tree with reduction factors h_1 and h_d obtained from data for the human lung¹⁹ and shown in Fig. 2; in this system, a 4% reduction in diameter doubles the resistance. The dotted line corresponds to an asymmetric tree with an average homothety ratio of $h = 0.85$; this system is slightly less sensitive to diameter reduction, but a 5% reduction in diameter still suffices to double the resistance.

surfaces around a branched duct system, and this provides the necessary exchange surface between air and blood¹⁵.

We can draw another general conclusion from the large sensitivity of the airway system to morphology. From a strictly physical point of view, minor differences among individuals can induce considerable differences in respiratory performance. This may explain why athletes are particularly sensitive to the effects of asthma or exercise-induced bronchospasm¹⁶. The higher performance of athletes resulting from increased aerobic capacity of muscle induced by training^{7,17,18} requires higher ventilation rates to ensure an adequate oxygen supply, but this is not accompanied by a commensurate training-induced adjustment of lung structure. The higher air-flow rates must be achieved in the given bronchial tree and thus airway geometry becomes critical. An exercise-induced narrowing of small airways results in a reduction of the homothety factor h_d , which causes the flow resistance to increase steeply. This would be particularly hazardous if the homothety ratio of relaxed airways was close to its critical value of 0.79, because athletes working at their limit might not have the reserve to increase the work of breathing that this would impose. The fact that there are athletes that are capable of high performance in spite of exercise-induced bronchospasm may indicate that the homothety factor of airway branching has a sufficient safety margin to allow them adequate ventilation of their pulmonary gas exchanger. A 'more optimal' design of the airway tree would be dangerous. □

Methods

The total tree resistance can be expressed exactly as:

$$R_N = R_0 \left[1 + \sum_{p=1}^N \frac{1}{2^p} \left(\frac{h_1}{h_d} \right)^p \right] = R_0 \frac{1 - r^{N+1}}{1 - r}$$

where $r = h_1/2h_d$. The tree resistance is then a nonlinear function of $1/h_d^4$.

Received 29 July; accepted 11 December 2003; doi:10.1038/nature02287.

- Mandelbrot, B. *The Fractal Geometry of Nature* (W. H. Freeman, San Francisco, CA, 1982).
- West, G. B., Brown, J. H. & Enquist, B. J. A general model for the origin of allometric scaling laws in biology. *Science* **276**, 122–126 (1997).
- Brown, J. H., West, G. B. & Enquist, B. J. *Scaling in Biology* (Oxford Univ. Press, Oxford, UK, 2000).
- Bejan, A. *Shape and Structure, From Engineering to Nature* (Cambridge Univ. Press, Cambridge, UK, 2000).
- Nelson, T. R. & Manchester, D. K. Modeling of lung morphogenesis using fractal geometries. *IEEE Trans. Med. Imaging* **7**, 321–327 (1988).
- West, B. J., Barghava, V. & Goldberger, A. L. Beyond the principle of similitude: renormalization in the bronchial tree. *J. Appl. Physiol.* **60**, 1089–1097 (1986).
- Weibel, E. R. *The Pathway for Oxygen* (Harvard Univ. Press, Cambridge, MA, 1984).
- Mauroy, B., Filoche, M., Andrade, J. S. & Sapoval, B. Interplay between geometry and flow distribution in an airway tree. *Phys. Rev. Lett.* **90**, 1–4 (2003).
- Hess, W. R. Das Prinzip des kleinsten Kraftverbrauchs im Dienste hämodynamischer Forschung. *Archiv Anat. Physiol.* **1914**, 1–62 (1914).
- Murray, C. D. The physiological principle of minimum work. I. The vascular system and the cost of blood. *Proc. Natl Acad. Sci. USA* **12**, 207–214 (1926).
- Weibel, E. R. in *The Lung: Scientific Foundations* 2nd edn Vol. 1 (eds Crystal, R. G., West, J. B., Weibel, E. R. & Barnes, P. J.) 1061–1071 (Lippincott-Raven, Philadelphia, PA, 1997).
- Que, C. L., Kenyon, C. M., Olivenstein, R., Macklem, P. T. & Maksym, G. N. Homeokinesis and short-term variability of human airway caliber. *J. Appl. Physiol.* **91**, 1131–1141 (2001).
- Sapoval, B. *Universalités et Fractales* (Flammarion, Paris, 1997).
- Kitaoka, H., Ryuji, T. & Suki, B. A three-dimensional model of the human airway tree. *J. Appl. Physiol.* **87**, 2207–2217 (1999).
- Sapoval, B., Filoche, M. & Weibel, E. R. Smaller is better—but not too small: a physical scale for the design of the mammalian pulmonary acinus. *Proc. Natl Acad. Sci. USA* **99**, 10411–10416 (2002).
- Wilber, R. L. et al. Incidence of exercise-induced bronchospasm in Olympic winter sport athletes. *Med. Sci. Sports Exerc.* **32**, 732–737 (2000).
- Weibel, E. R. *Symmorphosis* (Harvard Univ. Press, Cambridge, MA, 2000).
- Hoppeler, H. & Fluck, M. Plasticity of skeletal muscle mitochondria: structure and function. *Med. Sci. Sports Exerc.* **35**, 95–104 (2003).
- Weibel, E. R. *Morphometry of the Human Lung* (Springer, Berlin, 1963).

Acknowledgements The authors wish to thank J. M. Morel and M. Bernot for useful discussions.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to B.S. (Bernard.Sapoval@polytechnique.fr).

Susceptibility to leprosy is associated with *PARK2* and *PACRG*

Marcelo T. Mira^{1,2*}, Alexandre Alcaïs^{3*}, Nguyen Van Thuc⁴, Milton O. Moraes⁵, Celestino Di Flumeri¹, Vu Hong Thai⁴, Mai Chi Phuong⁴, Nguyen Thu Huong⁴, Nguyen Ngoc Ba⁴, Pham Xuan Khoa⁴, Euzenir N. Sarno⁵, Andrea Alter¹, Alexandre Montpetit⁶, Maria E. Moraes⁷, José R. Moraes⁷, Carole Doré⁶, Caroline J. Gallant¹, Pierre Lepage⁶, Andrei Verner⁶, Esther van de Vosse⁸, Thomas J. Hudson^{1,6}, Laurent Abel³ & Erwin Schurr¹

¹McGill Centre for the Study of Host Resistance and Departments of Human Genetics, Medicine and Biochemistry, McGill University, 1650 Cedar Avenue, Montreal, PQ H3G1A4, Canada

²Centro de Ciências Biológicas e da Saúde, Pontifícia Universidade Católica do Paraná, Rua Imaculada Conceição, 1155, CEP 80215-901, Curitiba, Paraná, Brasil

³Laboratoire de Génétique Humaine des Maladies Infectieuses, INSERM U.550, Faculté de Médecine Necker, Université de Paris René Descartes, 156 rue de Vaugirard, 75015 Paris, France

⁴Hospital for Dermato-Venereology, Nguyen Thong Street, District 3, Ho Chi Minh City, Vietnam

⁵Leprosy Laboratory, Tropical Medicine Department Oswaldo Cruz Institute, FIOCRUZ 20221-903 Rio de Janeiro, Brazil

⁶McGill University and Genome Québec Innovation Centre, 740 Docteur Penfield, Montreal, PQ H3A 1A4, Canada

⁷Laboratório de Imunogenética, Instituto Nacional do Cancer, Ministério da Saúde, 20230-130 Rio de Janeiro, RJ, Brazil

⁸Department of Infectious Diseases & Department of Immunohematology and Blood Transfusion, Leiden University Medical Center, Building 1, Albinusdreef 2, 2333 ZA Leiden, The Netherlands

* These authors contributed equally to this work

Leprosy is caused by *Mycobacterium leprae* and affects about 700,000 individuals each year¹. It has long been thought that leprosy has a strong genetic component², and recently we mapped a leprosy susceptibility locus to chromosome 6 region q25–q26 (ref. 3). Here we investigate this region further by using a systematic association scan of the chromosomal interval most likely to harbour this leprosy susceptibility locus. In 197 Vietnamese families we found a significant association between leprosy and 17 markers located in a block of approx. 80 kilobases overlapping the 5' regulatory region shared by the Parkinson's disease gene *PARK2* and the co-regulated gene *PACRG*. Possession of as few as two of the 17 risk alleles was highly predictive of leprosy. This was confirmed in a sample of 975 unrelated leprosy cases and controls from Brazil in whom the same alleles were strongly associated with leprosy. Variants in the regulatory region shared by *PARK2* and *PACRG* therefore act as common risk factors for leprosy.

The 90% confidence interval⁴ for placement of the chromosome 6q25–q26 leprosy susceptibility locus that we previously identified³ is defined by markers *D6S415* and *D6S1277*. The latter two markers are separated by 6.4 megabases (Mb) of chromosomal DNA on build 33 of the annotated human genome sequence (<http://www.ncbi.nlm.nih.gov>), and 31 genes with known or predicted function are located in this interval (Fig. 1a). By comparative genome sequencing and database searching, we developed a panel of 64 informative single-nucleotide polymorphisms (SNPs) (minor allele frequency more than 5%) with at least one SNP located close to or within each known gene (Fig. 1a, Supplementary Table 1). With these 64 markers, we conducted an association scan of the linkage peak interval among 197 Vietnamese families composed of two parents and one leprosy-affected child. Six markers were associated with leprosy susceptibility ($P < 0.05$), of which four, including the two most significant, were clustered in the promoter region of the *PARK2* and *PACRG* genes (Fig. 1a).

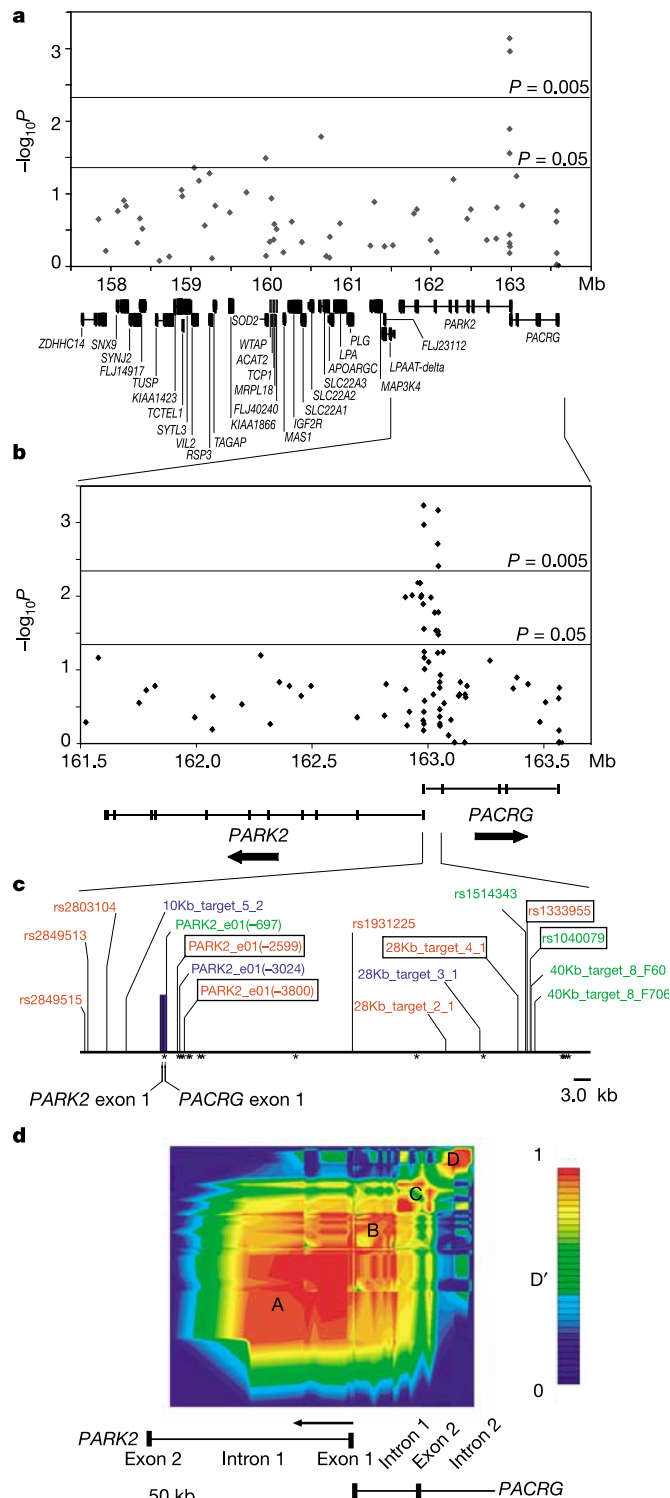


Figure 1 High-density SNP analysis. **a**, Evidence for association with leprosy of 64 SNPs located in the 90% confidence linkage interval is given as $-\log_{10}P$, and plotted against SNP position. The locations of 31 genes of this region (exons are denoted by vertical bars) are provided. **b**, $-\log_{10}P$ for 81 SNPs located within the *PARK2* and *PACRG* genes. **c**, Further expansion of an 80-kb region overlapping the *PARK2*/*PACRG* regulatory region. Boxed SNPs are associated with leprosy with $P < 0.005$. Asterisks indicate SNPs with $P > 0.05$. SNPs shown in red, green and blue display allele frequencies of 0.66/0.34, 0.84/0.16, and 0.51/0.49, respectively. **d**, LD between pairs of SNPs located in a 500-kb *PARK2*/*PACRG* region. Colour-coded D' values for pairs of SNPs are plotted with the GOLD program¹². Four regions of LD, marked A, B, C and D, are indicated.

In a further investigation of the role of the *PARK2* and/or *PACRG* genes in the control of leprosy susceptibility, we embarked on a systematic SNP discovery process based on a comparative sequencing of all *PARK2* and *PACRG* exons, their shared 5' regulatory region and both 3' untranslated regions. Although we detected several new SNPs, we did not find any new coding region SNPs in either gene. To increase SNP coverage we selected additional markers from public databases. Moreover, in regions with evidence for association, comparative genomic sequencing of small chromosomal segments was performed to obtain a higher marker density. By following this strategy, we identified 81 polymorphisms (mainly SNPs; Supplementary Table 1) with minor allele frequencies of at least 0.05 within the *PARK2* and *PACRG* genes (Fig. 1b). All markers were genotyped in the panel of 197 Vietnamese simplex families, and 19 SNPs clustered in the 5' *PARK2* and *PACRG* regulatory region were significantly associated ($P = 0.03$ – 0.0006) with leprosy (Fig. 1b and 1c).

We then constructed a linkage disequilibrium (LD) map spanning about 500 kilobases (kb) of the 5' region of *PARK2* and *PACRG*. Four LD regions could be distinguished (Fig. 1d): region A spanning about 200 kb of *PARK2* intron 1, region B overlapping the core promoter of *PARK2* and *PACRG* as well as about 80 kb of *PACRG* intron 1, and regions C and D in intron 1, exon 2 and intron 2 of the *PACRG* gene. There was extensive LD among SNPs between regions A and B. Most SNPs associated with leprosy localized to region B, but two associated SNPs (rs1893540 and rs2849531; $P = 0.01$) were found in region A more than 50 kb downstream of region B. The LD pattern clearly identified the 80-kb block overlapping the 5' regulatory regions of the *PARK2* and *PACRG* genes as the location of molecular variants controlling leprosy susceptibility (Fig. 1d). Among the 17 SNPs associated with leprosy in region B, we identified three groups of markers with respective common (minor) allele frequencies of ~ 0.66 (0.34), ~ 0.84 (0.16) and ~ 0.51 (0.49). Within each group, almost complete LD was observed (only two haplotypes accounted for more than 88% of all genotypes), and, as expected, strong LD was also observed between SNPs of the three groups (Fig. 1d).

Multivariate conditional logistic regression analysis indicated that only two 'tag' SNPs located within the first intron of *PACRG*, *PARK2*_e01(–2599) (belonging to group 0.66/0.34) and rs1040079 (belonging to group 0.84/0.16), were needed to capture all association information between the associated region and leprosy. Common allele 'T' of *PARK2*_e01(–2599) and rare allele 'C' of rs1040079 were independently associated with an increased risk of leprosy. In comparison with CC homozygotes, the odds ratio of

Table 1 Association with leprosy of genotypes defined from the three *PARK2*_e01(–2599)–rs1040079 haplotypes

<i>PARK2</i> (–2599)– rs1040079 genotypes	Frequency	Odds ratio	95% confidence interval	<i>P</i>
C–T	0.11	1.00		
T–T	0.32	3.15	1.29–7.71	0.012
T–T	0.26	3.68	1.41–9.62	0.008
T–C	0.10	5.35	1.92–14.89	0.001
T–C	0.17	5.72	2.10–15.59	0.0007
T–C	0.04	4.75	1.32–17.18	0.017

Accounting for dominance effect

C–T	C–T	0.11	1.00		
T–T	C–T/T–T	0.58	3.23	1.34–7.82	0.009
T–C	T–C/T–T/C–T	0.31	5.28	2.06–13.55	0.0005

Frequencies have been estimated from non-transmitted parental haplotypes. Odds ratios were computed by means of conditional logistic regression. The overall comparison showed a significant association with leprosy, accounting ($P = 0.0002$) or not ($P = 0.006$) for dominance effect. Likelihood ratio tests showed that the genotypic model including three risk levels denoted as 'accounting for dominance effect' (that is, T–T haplotype is dominant over C–T, and T–C haplotype is dominant over both T–T and C–T) was not rejected against the full genotypic model with six risk levels ($\chi^2 = 0.54$, 3 degrees of freedom, not significant), whereas the genotypic model with only two risk levels (T–T and T–C haplotypes have the same dominant effect over C–T) is rejected against the genotypic model with three risk levels ($\chi^2 = 4.21$, 1 degree of freedom, $P = 0.04$). This indicates that the best-fitting genotypic model was the one denoted in the table as 'accounting for dominance effect'.

Table 2 Association between leprosy and DNA polymorphisms of the *PARK2/PACRG* region

Marker name	Position	Vietnamese sample			Brazilian sample		
		Risk allele	<i>q</i>	<i>P</i>	Risk allele	<i>q</i>	<i>P</i> (<i>P</i> _{GC})
rs2803104	162972608	A	0.68	0.011	A	0.58	n.s.
10Kb_target_5_2	162976030	T	0.51	0.013	T	0.84	0.008 (0.019)
PARK2_e01(−697)	162982925	G	0.17	0.013	G	0.38	0.0002 (0.001)
PARK2_e01(−2599)	162984827	T	0.67	0.0006	T	0.61	0.0006 (0.003)
PARK2_e01(−3024)	162985252	C	0.52	0.029	C	0.83	n.s.
PARK2_e01(−3800)	162986028	G	0.66	0.001	G	0.61	0.003 (0.009)
28Kb_target_2_1	163032514	T	0.67	0.017	T	0.58	n.s.
28Kb_target_4_1	163045217	A	0.66	0.002	A	0.53	0.0009 (0.003)
rs1514343	163046511	T	0.17	0.03	T	0.23	0.023 (0.045)
rs1333955	163046882	C	0.66	0.0007	C	0.52	0.016 (0.034)
rs1040079	163047455	C	0.17	0.004	C	0.29	0.0002 (0.001)
40Kb_target_8_F60	163047538	A	0.16	0.034	A	0.24	0.015 (0.032)
40Kb_target_8_F706	163048194	G	0.16	0.017	G	0.23	n.s.

The position for each SNP is given as the nucleotide number on build 33 of the human chromosome 6 sequence map. *q* denotes the population frequency of the risk allele and was estimated from non-transmitted parental alleles in the Vietnamese sample and among unaffected controls in the Brazilian sample. *P*_{GC} denotes the robust *P* value obtained by using 24 genomic controls as proposed in ref. 18. n.s., not significant at the 0.05 level.

leprosy for carriers of at least one T allele of PARK2_e01(−2599) was 2.78 (95% confidence interval 1.3–5.95), and the odds ratio for carriers of at least one C allele of rs1040079 versus TT homozygotes was 1.80 (1.15–2.81). Owing to strong LD, the two tag SNPs defined only three haplotypes among the Vietnamese families, T–C (containing the T risk allele for PARK2_e01(−2599) and the C risk allele for rs1040079), T–T, and C–T (carrying the two protective alleles); the C–C haplotype was not observed. We therefore used haplotypic information to investigate further the association of the two tag SNPs with leprosy.

The three two-marker SNP haplotypes defined six genotypes that were identified unambiguously in the Vietnamese families and showed strong association with leprosy (Table 1). As expected, subjects homozygous for the C–T haplotype had the lowest risk of leprosy and were subsequently used as the baseline risk population. Formal conditional logistic analysis showed that the T–T haplotype was dominant over C–T, and that the T–C haplotype was dominant over T–T and C–T (Table 1). In comparison with C–T homozygotes, the odds ratio of leprosy for individuals carrying at least one T–C haplotype was 5.28 (2.06–13.55), and that for carriers of at least one T–T haplotype (without T–C haplotype) was 3.23 (1.34–7.82). These results strongly implicate specific haplotypes of the 80-kb LD block (region B in Fig. 1d) in leprosy risk in Vietnamese patients. To identify all the polymorphisms in LD with the risk haplotypes, we sequenced more than 95% of the 80-kb haplotype block for at least two leprosy-affected members of each of the five genotypic

risk groups and detected 185 new polymorphisms, mainly SNPs (Supplementary Table 2). Although the number of chromosomes by this resequencing project is limited to 24, we observed co-occurrences of 51% of the new variants with alleles of known SNPs defining the three previously identified groups of markers of region B. The remaining new polymorphisms were either infrequent (that is, one or two observances; 36% of new polymorphism) or were not clearly correlated with any of the risk haplotypes (13%). The high comparative sequencing coverage suggests that all major genetic risk variants in the critical 80-kb interval have been identified.

To confirm these results we analysed an independent sample of 975 unrelated individuals (587 cases affected by leprosy and 388 unaffected controls) from Rio de Janeiro, Brazil. Of the 17 markers in region B (Fig. 1c, d) that were significantly associated with leprosy in Vietnamese families, we selected 15 that were assayed for their allelic composition in the Brazilian sample; valid genotypes were obtained for 13 markers (Table 2). Of these 13 SNPs, 9 were significantly associated with leprosy in the Brazilian sample, with the same alleles being associated with an increased risk of leprosy in the two populations (Table 2). Interestingly, PARK2_e01(−2599) and rs1040079 were among the three most significantly associated SNPs (*P* < 0.0006, *P*_{GC} < 0.003) in the Brazilian population.

LD studies of the 13 SNPs in the Brazilian population showed patterns similar to those observed in Vietnam with generally weaker LD. Additionally, no familial inheritance data were available,

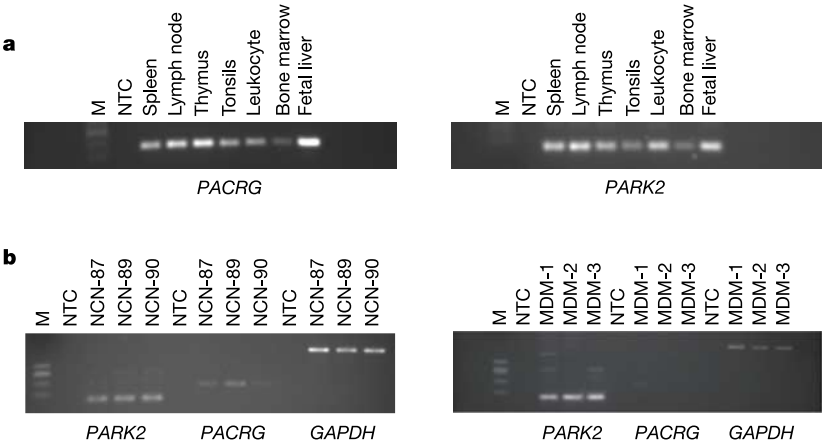


Figure 2 Expression of *PARK2* and *PACRG*. **a**, Both genes are expressed in all immune tissues, with lower expression in bone marrow. **b**, Expression in three primary human Schwann cell lines (NCN 88, NCN 89 and NCN 90) and monocyte-derived macrophages (MDM-1, MDM-2 and MDM-3) obtained from three donors. Both genes are expressed by

Schwann cells, whereas in monocyte-derived macrophages *PARK2* is more readily detectable. However, with increased polymerase chain reaction cycles, a band can also be detected with *PACRG*-specific primers in monocyte-derived macrophage cDNA (not shown). M, molecular mass marker; NTC, non-template control.

making it impossible to reconstruct marker haplotypes unambiguously. To test for replication of the haplotype results observed in Vietnam we therefore introduced a binary variable, a 'meta-SNP', combining the information at SNPs PARK2_e01(−2599) and rs1040079. This meta-SNP takes the value 1 for subjects carrying at least one risk allele at both of these SNPs, and the value 0 otherwise. A multivariate stepwise logistic regression analysis including the meta-SNP and all SNPs associated with leprosy showed that the meta-SNP captured all association information between the PARK2/PACRG 5' regulatory region and leprosy in the Brazilian patients ($P < 0.0000005$, $P_{GC} < 0.00002$; odds ratio 2.21 (1.61–3.01)), in perfect analogy with the situation in the Vietnamese families. Overall, these association studies performed in two distinct populations identify the regulatory region of PACRG and PARK2 as a major risk factor for leprosy.

Next we tested the expression pattern of PARK2 and PACRG in selected human tissues and cell types. Both genes are widely expressed in human tissues⁵, including immune tissues (Fig. 2a), with loosely correlated expression levels. This observation suggests that, in addition to the common bi-directional promoter⁵, unknown gene-specific transcriptional activators are involved in regulating cell-specific and tissue-specific gene expression. Further, we tested the expression pattern of PARK2 and PACRG in human Schwann cells and macrophages, which are the primary host cells of *M. leprae*⁶. PARK2 is expressed by primary Schwann cell explants and by monocyte-derived macrophages, whereas PACRG is strongly expressed in primary Schwann cells and very weakly by monocyte-derived macrophages (Fig. 2b). Expression of PARK2/PACRG by the *M. leprae* host cells supports the results of the association studies.

Here we report the first successful identification by positional cloning of genetic variants with an important function in a common human infectious disease. We obtained these results in two populations with very different ethnic backgrounds, thus identifying specific variants of the 5' region of PACRG and PARK2 as worldwide risk factors for leprosy. Mutations in the PARK2 gene have previously been shown to be the cause of autosomal recessive early-onset Parkinson's disease⁷. Functionally, PARK2 is a ubiquitination E3 ligase that is directly involved in the delivery of polyubiquitinated proteins to the proteasome complex⁸. The function of PACRG is unknown but also has been linked to the ubiquitin–proteasome system^{5,9}. Our genetic findings combined with the expression of PARK2/PACRG in macrophages and Schwann cells therefore point strongly to an important function of ubiquitin-mediated proteolysis—a biochemical pathway that has so far received little attention in the study of leprosy pathogenesis—for the control of *M. leprae* in the human host. □

Methods

Subjects

A collection of 197 simplex leprosy families of Vietnamese ethnicity was enrolled as described previously³. For the replication study, 587 Brazilian leprosy patients were recruited from the Leprosy Out-Patient Unit at the Oswaldo Cruz Institute, Rio de Janeiro, Brazil. In addition, 388 control subjects with no history of chronic infectious, autoimmune or inflammatory disease were recruited from the same geographic area of the city of Rio de Janeiro. There was no significant difference in the proportion of the three main ethnic groups, African-Brazilians, Euro-Brazilians and Mestizos, between cases and controls. The study was approved by institutional review boards and health authorities in Ho Chi Minh City, Vietnam; the Oswaldo Cruz Institute, Rio de Janeiro, Brazil; and the McGill University Health Centre, Montreal, Canada.

Phenotype definition

Diagnosis and clinical subtype definition for the Vietnamese patients have been described³. Brazilian cases were classified on the basis of clinical and histological criteria¹⁰, and also included a mixture of paucibacillary (38%) and multibacillary (62%) cases. The phenotype analysed for both Vietnamese and Brazilian patients was leprosy *per se*; that is, leprosy independent of specific clinical presentations.

SNP identification

The identification of informative intragenic SNPs was done by sequencing 12 selected Vietnamese individuals for all PARK2 and PACRG exons, as well as 3.0 kb upstream of

exon 1 of PARK2, 1.0 kb upstream of exon 1 of PACRG, and both 3' untranslated regions. SNPs identified in the above regions but not included in the SNP database were named according to their position in GenBank accession number AB009973 relative to PARK2 exon 1. Additional SNPs were selected from the SNP Consortium (<http://snp.cshl.org/>) and the dbSNP-NCBI (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=snp>) public databases that were mapped to build 33 of the NCBI human genome assembly. If necessary, roughly 1-kb chromosomal DNA segments uniformly distributed along regions of low marker coverage were sequenced from 12 selected individuals, and sequences were scanned for additional SNPs (comparative sequence skimming). SNPs identified by comparative sequence skimming of the extended shared promoter region were named with the sequence target name followed by a unique identifier. For the genomic controls, all Brazilian DNA samples were genotyped with 24 SNPs (minor allele frequency more than 0.16) distributed randomly throughout the genome.

Genotyping

All individuals from both the Vietnamese and Brazilian collections were genotyped by one of the following methods: direct comparative sequencing, SSLP analysis, fluorescent polarization assays, and single-base-pair extension on the Ultra-High Throughput Orchid platform (Supplementary Table 1). In the Vietnamese sample, genotyping errors were estimated from apparent non-mendelian transmissions of alleles, which were consistently found to be less than 1%. To assess genotype concordance in the Brazilian sample, three selected markers (PARK2_e01(−2599), PARK2_e01(−697) and 28kb_target_4_1) were double genotyped by both the Ultra-High Throughput Orchid and fluorescent polarization methods. Genotypes for discordant samples were obtained by direct sequencing.

Gene expression analysis

For expression analysis by gene amplification, we used Human Multiple Tissue and Human Immune System cDNA panels from ClonTech (BD Biosciences, Palo Alto, California, USA) and complementary DNA constructed from RNA obtained with a Qiagen RNA extraction kit from monocyte-derived macrophages and primary Schwann cell explants. All amplifications were performed under identical conditions at an annealing temperature of 56 °C, with 25 cycles of amplification for GAPDH and 38 cycles of amplification for PARK2 and PACRG.

Statistical methods

To estimate pairwise LD between SNPs we measured D' from parental data in the Vietnamese sample and from unaffected controls in the Brazilian sample, using the algorithm proposed in ref. 11 as implemented in the GOLD software¹². All LD patterns were depicted with the GOLD software. Two SNP haplotype frequencies were estimated with the HAPLOTYPYER software¹³.

Because the Vietnamese sample consisted of trios with no missing parental data, family-based association study was performed mainly with a classical transmission disequilibrium test as implemented in the software FBAT¹⁴ and using the empirical variance–covariance estimator previously advocated¹⁵. Population allelic frequencies were estimated on untransmitted parental alleles as proposed in ref. 16. Alleles showing some evidence for association were also analysed by conditional logistic regression as described in ref. 17. This approach permitted providing odds-ratio estimates and performing multivariate stepwise regression. Conditional logistic analysis was performed with the PHREG procedure implemented in the SAS software v. 8.1 (SAS Institute, Cary, North Carolina, USA).

Population-based association study in the Brazilian sample was performed with classical multivariate logistic regression techniques as implemented in the LOGISTIC procedure of the SAS software v. 8.1. All analyses were performed by adjusting on gender, because gender is a classical risk factor associated with leprosy. To account for possible cryptic population stratification, we also used the genomic control method¹⁸ to obtain robust P -values denoted as P_{GC} (see Supplementary Methods).

Received 16 September 2003; accepted 7 January 2004; doi:10.1038/nature02326.

Published online 25 January 2004.

- WHO, Leprosy. Global situation. *Wkly Epidemiol. Rec.* **77**, 1–8 (2002).
- Casanova, J. L. & Abel, L. Genetic dissection of immunity to mycobacteria: The human model. *Annu. Rev. Immunol.* **20**, 581–620 (2002).
- Mira, M. T. et al. Chromosome 6q25 is linked to susceptibility to leprosy in a Vietnamese population. *Nature Genet.* **33**, 412–415 (2003).
- Dupuis, J. & Siegmund, D. Statistical methods for mapping quantitative trait loci from a dense set of markers. *Genetics* **151**, 373–386 (1999).
- West, A. B., Lockhart, P. J., O'Farrell, C. & Farrer, M. J. Identification of a novel gene linked to parkin via a bi-directional promoter. *J. Mol. Biol.* **326**, 11–19 (2003).
- Jacobson, R. R. & Krahenbuhl, J. L. Leprosy. *Lancet* **353**, 655–660 (1999).
- Kitada, T. et al. Mutations in the parkin gene cause autosomal recessive juvenile parkinsonism. *Nature* **392**, 605–608 (1998).
- Giasson, B. I. & Lee, V. M. Parkin and the molecular pathways of Parkinson's disease. *Neuron* **31**, 885–888 (2001).
- Imai, Y. et al. A product of the human gene adjacent to parkin is a component of Lewy bodies and suppresses Pael receptor-induced cell death. *J. Biol. Chem.* **278**, 51901–51910 (2003).
- Ridley, D. S. & Jopling, W. H. Classification of leprosy according to immunity. A five-group system. *Int. J. Lepr. Other Mycobact. Dis.* **34**, 255–273 (1966).
- Excoffier, L. & Slatkin, M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.* **12**, 921–927 (1995).
- Abecasis, G. R. & Cookson, W. O. GOLD—graphical overview of linkage disequilibrium. *Bioinformatics* **16**, 182–183 (2000).
- Niu, T., Qin, Z. S., Xu, X. & Liu, J. S. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Am. J. Hum. Genet.* **70**, 157–169 (2002).

14. Horvath, S., Xu, X. & Laird, N. M. The family based association test method: Strategies for studying general genotype-phenotype associations. *Eur. J. Hum. Genet.* **9**, 301–306 (2001).
15. Lake, S. L., Blacker, D. & Laird, N. M. Family-based tests of association in the presence of linkage. *Am. J. Hum. Genet.* **67**, 1515–1525 (2000).
16. Thomson, G. Mapping disease genes: Family-based association studies. *Am. J. Hum. Genet.* **57**, 487–498 (1995).
17. Schaid, D. J. & Rowland, C. Use of parents, sibs, and unrelated controls for detection of associations between genetic markers and disease. *Am. J. Hum. Genet.* **63**, 1492–1506 (1998).
18. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank all patients, their families and population controls that participated in this study; J. L. Casanova for support, suggestions and critical reading of the manuscript; T. H. M. Ottenhoff for his help and support in the Schwann cell work; and A. Sammak, C. Darmond-Zwaig, Y. Renaud, M. Girard, N. Martin, A. Villeneuve, D. Frappier and A. Verville for technical assistance. This study was supported by a grant from the Canadian Institutes of Health Research to E.S. The genotyping core at the McGill University and Genome Québec Innovation Centre is supported by the Canadian Genetic Diseases Network and Genome Québec, and the Schwann cell work in Leiden is supported by the Netherlands Leprosy Foundation. M.T.M. is supported by a graduate fellowship of the Brazilian Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). A.A. and L.A. are supported in part by grants from Fondation Banque Nationale de Paris-Paribas, Fondation pour la Recherche Médicale and Fondation Schlumberger. T.J.H. and E.S. are Canadian Institutes of Health Research Investigators.

Authors' contributions L.A. and E.S. share senior authorship.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to L.A. (abel@necker.fr) or E.S. (erwin@igloo.epi.mcgill.ca).

Identification of an angiogenic factor that when mutated causes susceptibility to Klippel–Trenaunay syndrome

Xiao-Li Tian^{1*}, Rajkumar Kadaba^{1*}, Sun-Ah You^{1*}, Mugen Liu^{1,2*}, Ayse Anil Timur¹, Lin Yang³, Qiuyun Chen⁴, Przemyslaw Szafranski⁵, Shaoqi Rao¹, Ling Wu¹, David E. Housman⁶, Paul E. DiCorleto³, David J. Driscoll⁷, Julian Borrow^{6†} & Qing Wang¹

¹Center for Molecular Genetics, Department of Molecular Cardiology, Lerner Research Institute, and Center for Cardiovascular Genetics, Department of Cardiovascular Medicine, The Cleveland Clinic Foundation, and Department of Molecular Medicine, Cleveland Clinic Lerner College of Medicine of Case Western Reserve University, Cleveland, Ohio 44195, USA

²Institute of Genetics, Fudan University, Shanghai 200433, China

³Department of Cell Biology, and ⁴Cole Eye Institute, The Cleveland Clinic Foundation, Cleveland, Ohio 44195, USA

⁵Department of Pathology, Baylor College of Medicine, Houston, Texas 77030, USA

⁶Center for Cancer Research, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

⁷Division of Pediatric Cardiology, Mayo Clinic, Rochester, Minnesota 55905, USA

* These authors contributed equally to this work

† Present address: Cancer Research UK, Leukemia Molecular Genetics, Paterson Institute for Cancer Research, Manchester, M20 4BX, UK

Angiogenic factors are critical to the initiation of angiogenesis and maintenance of the vascular network¹. Here we use human genetics as an approach to identify an angiogenic factor, VG5Q, and further define two genetic defects of VG5Q in patients with the vascular disease Klippel–Trenaunay syndrome (KTS)^{2,3}. One mutation is chromosomal translocation t(5;11), which increases VG5Q transcription. The second is mutation E133K identified in

five KTS patients, but not in 200 matched controls. VG5Q protein acts as a potent angiogenic factor in promoting angiogenesis, and suppression of VG5Q expression inhibits vessel formation. E133K is a functional mutation that substantially enhances the angiogenic effect of VG5Q. VG5Q shows strong expression in blood vessels and is secreted as vessel formation is initiated. VG5Q can bind to endothelial cells and promote cell proliferation, suggesting that it may act in an autocrine fashion. We also demonstrate a direct interaction of VG5Q with another secreted angiogenic factor, TWEAK (also known as TNFSF12)^{4,5}. These results define VG5Q as an angiogenic factor, establish VG5Q as a susceptibility gene for KTS, and show that increased angiogenesis is a molecular pathogenic mechanism of KTS.

Angiogenesis has an essential role in pathological conditions such as cancer and various ischaemic and inflammatory diseases^{1,6–9}. KTS (OMIM number 149000) is a congenital vascular disease characterized by malformations of capillary (98% of KTS patients), venous (72%) and lymphatic (11%) vessels, and bony and soft tissue hypertrophy (67%)^{2,3}. KTS is commonly sporadic and its aetiology is unknown. To identify a susceptibility gene for KTS, we characterized translocation t(5;11)(q13.3;p15.1), which is associated with KTS¹⁰. Polymerase chain reaction (PCR) analysis of somatic cell hybrids containing only the derivative chromosome 5 or the derivative chromosome 11 defined the precise locations of two translocation breakpoints (Fig. 1a; see also Supplementary Fig. 1). A novel gene, VG5Q, was identified at the 5q13.3 translocation breakpoint. The 5q13.3 translocation breakpoint is located in the promoter of VG5Q (Supplementary Fig. 2), and significantly affects the expression of VG5Q (see below). The full-length VG5Q complementary DNA (4,049 base pairs (bp)) contains a long open reading frame that spans 2,145 bp and encodes a novel protein of 714 amino acids with a forkhead-associated (FHA) domain¹¹ and a G-patch domain¹² (amino acids 435–508 and 619–663, respectively; Fig. 1b). No genes were identified in the regions 50 kilobases (kb) on either side of the chromosome 11p15.1 breakpoint (Supplementary Fig. 1b).

To provide further evidence that VG5Q is a susceptibility gene for KTS, we performed a case-control study with 130 unrelated KTS cases and 200 unrelated matched controls. Genetic case-control association studies are a powerful alternative to family-based designs as a method of choice for identifying susceptibility genes for complex diseases as well as sporadic disorders such as KTS¹³. Explicit attempts were made to match the cases with controls for age, gender and ethnicity. Genotyping with population-specific markers^{13–15} was used to confirm that our cases matched controls. Allelic frequencies of the markers genotyped, including a single-nucleotide polymorphism P698T in VG5Q, are statistically similar in cases and controls ($P > 0.05$), and they all resemble the frequencies of the Caucasian population (Supplementary Table 1). Mutation analysis of VG5Q using the primers listed in Supplementary Table 2 revealed a non-conservative mutation E133K in five independent KTS patients, which resulted in substitution of a negatively charged glutamine residue by a positively charged lysine residue (Fig. 1c, d). All five patients are heterozygous for mutation E133K. Identification of VG5Q mutation E133K in five independent KTS patients suggests that E133 is a mutational hotspot. As E133K occurs within the context of a CpG dimer, it is expected to be a potential hotspot for mutations¹⁶. Mutation E133K was not detected in 200 matched controls. A statistically significant association was thus established between E133K of VG5Q and KTS ($P = 0.009$). This conclusion was also validated using a new genetic-based association study, Genomic Control (estimated λ by mean = max (0.66, 1))¹⁴. These results suggest that VG5Q is a susceptibility gene for KTS and implicate VG5Q in the pathogenesis of the disorder.

VG5Q was clearly detected in cells relevant to KTS, including endothelial cells, vascular smooth muscle cells (VSMCs) and osteoblasts (MG-63) (Fig. 2A–D). Northern blot analysis revealed

a single 4.5-kb transcript in human umbilical vein endothelial cells (HUVECs) and human microvascular endothelial cells (HMVECs) (Fig. 2A). Western blot analysis with an anti-VG5Q antibody recognized a predicted 84-kDa protein present in HMVECs and HUVECs (Fig. 2B). VG5Q was also expressed in human heart fibroblast (HHF) and ovarian cancer cells (OV-3), but low expression was detected in kidney cancer cells (RP-45), HeLa cells, and bladder cancer cells (T-24) (Fig. 2D). VG5Q was ubiquitously expressed in the human tissues examined (Fig. 2C). Co-immunostaining showed that VG5Q protein co-localized with an endothelial-specific marker, CD31, as well as a VSMC-specific marker, smooth muscle cell α -actin, in blood vessels embedded in various mouse tissues including heart, kidney, tail and limb (Fig. 2E; see also Supplementary Fig. 3).

VG5Q protein was detected mostly in the cytoplasm and around the nuclei of HMVECs (Fig. 2F). In an *in vitro* model of angiogenesis where endothelial cells were plated onto matrigel, VG5Q protein began to redistribute by moving towards the cell periphery and was also detected outside the cell (Fig. 2F, d, f, h). The dynamic re-distribution of VG5Q at this stage resembles the secretion pattern of other secreted proteins¹⁷. These results suggest that VG5Q may be secreted when endothelial cell tube formation is initiated. To confirm that VG5Q is secreted during angiogenesis, a competitive

enzyme-linked immunosorbent assay (ELISA) was carried out. Purified VG5Q results in reduced absorbance in the competitive ELISA assay (Fig. 2G). Similarly, the media from matrigel culture also led to significantly reduced absorbance in the competitive ELISA assay compared with the media from plastic culture and blank media (media from matrigel-coated plates without cells) ($P = 0.009$), indicating that the matrigel culture media contained secreted VG5Q. Metabolic labelling with [³⁵S]methionine/[³⁵S]cysteine revealed the presence of VG5Q in the matrigel media of HUVECs transfected with a VG5Q expression construct, but not with vector control (Fig. 2H). These results indicate that angiogenesis accompanies secretion of VG5Q protein. The molecular mechanisms for trafficking and secretion of VG5Q remain to be established; however, it may be released via a non-classical secretory pathway similar to the angiogenic factor FGF-2 (ref. 18).

VG5Q functions as an angiogenic factor. Similar to VEGF, purified wild-type VG5Q protein promoted strong angiogenesis in a chick chorioallantoic membrane (CAM) assay, as shown by the newly formed radiated vessels on the CAM with VG5Q (Fig. 3A). These results suggest that VG5Q is a potent angiogenic factor. Consistent with this finding, using the RNA interference technique, short interfering RNAs (siRNAs) against VG5Q reduced VG5Q expression at both messenger RNA and protein levels, and inhibited

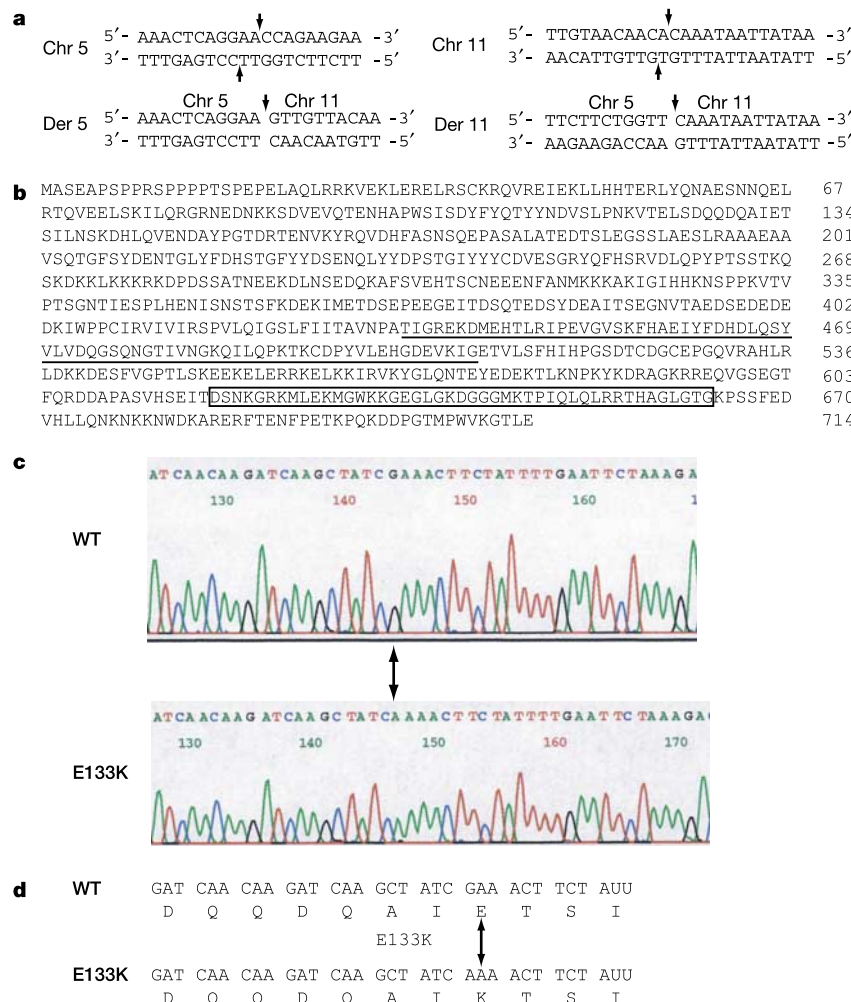


Figure 1 Positional cloning of the *VG5Q* gene. **a**, Definition of chromosome breakpoints involved in translocation t(5;11)(q13.3;q15.1) associated with KTS. Chr 5/Chr 11, normal chromosomes; Der 5/Der 11, derivative chromosomes. **b**, The amino acid sequences of human *VG5Q* protein. The FHA domain (435–508) and a G-patch domain (619–663) are

indicated. The estimated pI of VG5Q is 5.2. **c**, **d**, Identification of VG5Q mutation E133K in five independent KTS patients. **c**, DNA sequence analysis (G to A substitution at codon 133). **d**, Schematic representation of mutation E133K (substitution of a glutamic acid residue by a lysine residue).

tube formation by HMVECs and HUVECs that were plated onto matrigel (Fig. 3B). Similar findings were observed with an antisense oligonucleotide specific to VG5Q (5'-ATCACAAAAA-TAGTCCCC-3').

Purified VG5Q was able to bind to cultured endothelial cells (Fig. 3C) in calcein AM-based cell adhesion assays. Significantly higher calcein fluorescence (cells bound) was detected with VG5Q protein than control bovine serum albumin (BSA) (Fig. 3C). As VG5Q binds to the surface of endothelial cells and triggers endothelial cell proliferation (see below), it may act on endothelial cells in an autocrine fashion. However, a paracrine action of VG5Q cannot be ruled out as it is also expressed in VSMCs. It remains to be determined whether VG5Q binds to endothelial cells through a receptor. If such a receptor exists it may be protein in nature, as trypsinized endothelial cells, in contrast with endothelial cells collected by cell dissociation buffer, no longer adhered to VG5Q.

Functional characterization of two KTS-associated genetic mutations—one translocation t(5:11) found in one patient and one missense mutation E133K in 5 of 130 patients—validated VG5Q as a susceptibility gene for KTS. First, VG5Q promoter with the translocation junction fragment showed threefold higher expression than the wild-type VG5Q promoter (Fig. 3D, construct (3) versus construct (2)). These data suggest that the t(5:11) KTS translocation is a functional genetic defect that can lead to over-expression of VG5Q, which may result in increased angiogenesis. Second, significant differences in angiogenesis were observed between wild-type and mutant VG5Q with the E133K substitution ($P < 0.05$) (Fig. 3A, j). Mutant VG5Q protein acted as a more

potent angiogenic factor than the wild-type protein (wild type versus mutant: Fig. 3A, c versus d at a concentration of $37.5 \text{ ng } \mu\text{l}^{-1}$; e versus f at $75 \text{ ng } \mu\text{l}^{-1}$; g versus h at $150 \text{ ng } \mu\text{l}^{-1}$; summarized in panel j, $P < 0.05$). These results demonstrate that mutation E133K of VG5Q is a functional mutation that acts by a gain-of-function mechanism (increased angiogenesis). Mutation E133K is located within a putative phosphorylation site for glycogen synthase kinase 3 (I132-E133-T134-S135-I136-L137-N138-S139), and is only one amino-acid residue away from a putative phosphorylation site for protein kinase casein kinase 1 (S135-I136-L137-N138-S139-K140-D141-H142). The change of the phosphorylation state of the endothelial-cell-specific receptor tyrosine kinase TIE2 can cause vascular venous malformations (one of the clinical features manifested in KTS). So far only two mutations have been identified in TIE2, and both show ligand-independent hyperphosphorylation and result in constitutively active TIE2 (ref. 19). Thus, we speculate that mutation E133K may affect the phosphorylation state of VG5Q, resulting in increased angiogenesis. Histological analysis of subcutaneous veins in 33 KTS patients revealed an increase in the number of small venules (capillary veins) in affected KTS tissues²⁰. Our 'increased angiogenesis' theory may explain the histological features of KTS. Increased angiogenesis will lead to an increased number of blood vessels, as observed in affected tissues from KTS patients.

To elucidate the molecular mechanism by which VG5Q promotes angiogenesis, we isolated protein factors that associate with VG5Q using the yeast two-hybrid system with VG5Q as the 'bait'. Sequence analysis showed that one isolated cDNA encoded the carboxy-

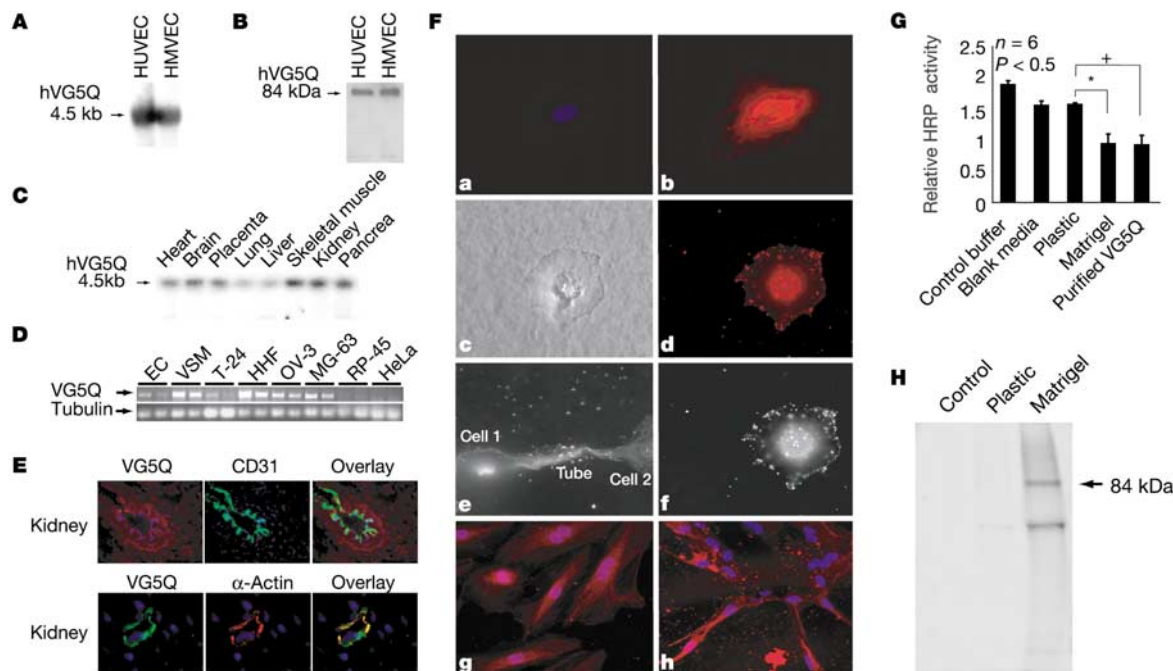


Figure 2 Expression profile of VG5Q and dynamic redistribution and secretion of VG5Q protein during angiogenesis. **A**, **B**, VG5Q expression in HMVECs and HUVECs determined by northern (**A**) and western blot analyses (**B**). **C**, Tissue expression pattern of VG5Q determined by northern blot analysis. **D**, Expression of VG5Q in different cell lines determined by RT-PCR. Tubulin, internal control. **E**, Immunostaining of mouse kidney for VG5Q protein expression. The nucleus was stained with DAPI (blue signal). CD31, endothelial cell marker; α -actin, VSMC marker. **F**, HMVECs were cultured on plastic (**a**, **b**, **g**) and matrigel (**c**–**f**, **h**). VG5Q protein, red; nucleus, blue (DAPI). **c**, **d**, **f**, One hour on matrigel; **e**, **g**, 4 h on matrigel; **h**, 24 h on matrigel. **G**, Competitive ELISA analysis to show

that VG5Q is secreted. Relative HRP activity, absorbance reading of the wells. An asterisk indicates statistical significance ($P < 0.05$). **H**, Detection of VG5Q in the media by metabolic labelling and immunoprecipitation, indicating that cells secrete VG5Q into the media. Control, HUVECs transfected with pcDNA3.1; plastic and matrigel, HUVECs transfected with pcDNA3.1-VG5Q and plated on either plastic or matrigel plates, respectively. The 84-kDa band indicated by an arrow represents VG5Q. The nature of the small 45-kDa band is unknown, but may be a cleaved VG5Q product. Faint signal was also observed in the media from cells cultured on plastic dishes. This may suggest weak secretion of VG5Q under this condition.

terminal domain of TWEAK (amino acid residues 136–249), a member of the tumour-necrosis factor (TNF) superfamily that induces angiogenesis *in vivo*²¹. The direct physical interaction between VG5Q and TWEAK was demonstrated using glutathione S-transferase (GST) pull-down assays with GST–TWEAK protein and *in-vitro*-translated ³⁵S-labelled VG5Q (Fig. 4a). In co-immunoprecipitation assays, the anti-VG5Q antibody specifically precipitated a protein recognized by an anti-TWEAK antibody, validating

the interaction between VG5Q and TWEAK *in vivo* (Fig. 4b). Co-immunostaining showed co-localization of the two proteins around the nuclei in HUVECs cultured on plastic dishes (Fig. 4c). In HUVECs initiating endothelial tube formation on matrigel, VG5Q and TWEAK moved to the cell surface (Fig. 4c). Together, these results suggest that VG5Q may promote angiogenesis by interacting with TWEAK.

TWEAK binds to its receptor, fibroblast-growth-factor-inducible

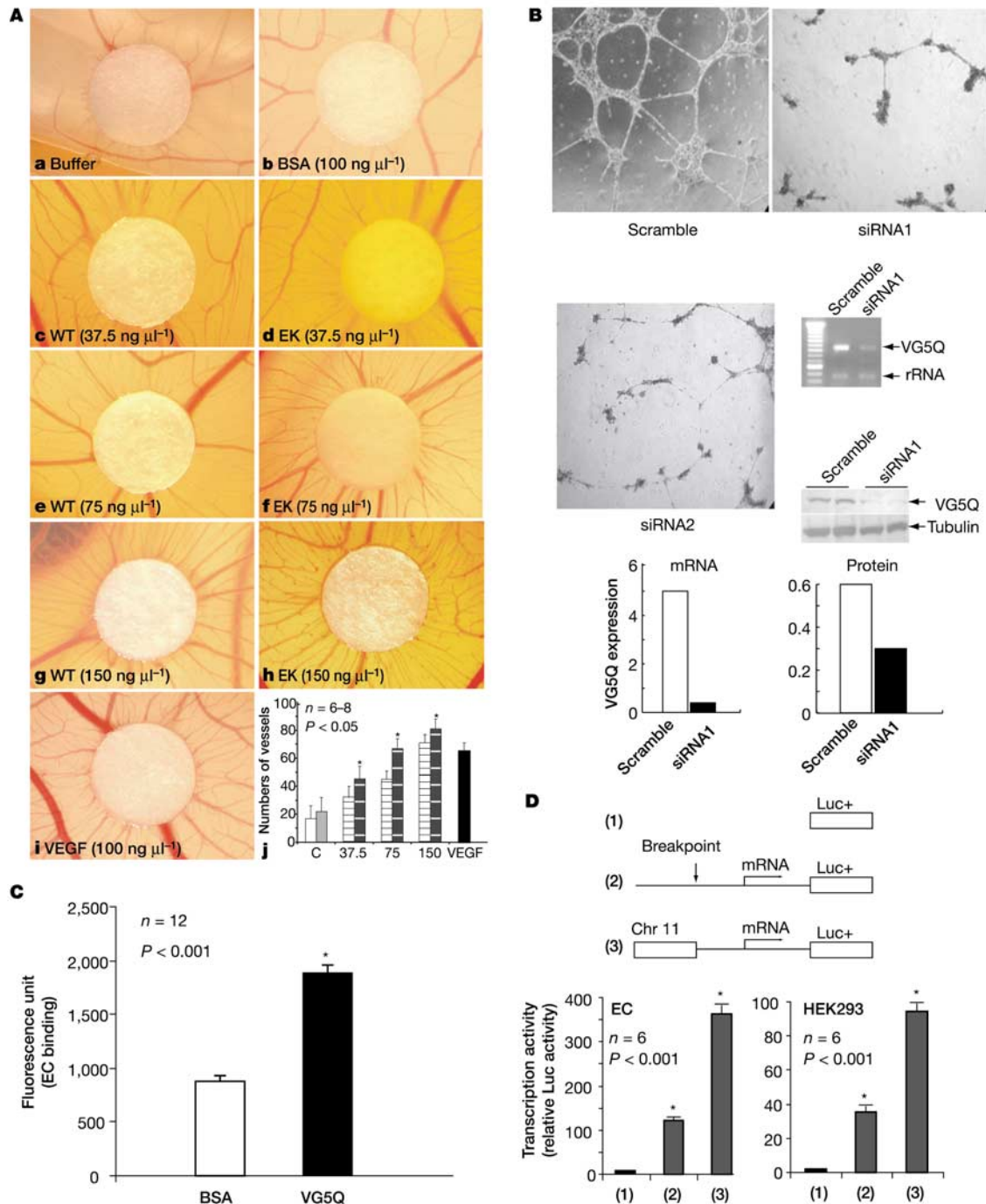


Figure 3 VG5Q is an angiogenic factor and both VG5Q E133K and KTS translocation t(5;11) are functional mutations. **A**, Angiogenesis promoted by VG5Q protein was determined by CAM assays. EK, VG5Q with mutation E133K. **B**, Effect of siRNA against VG5Q on endothelial tube formation. Scramble, control siRNA. Suppression of VG5Q expression by siRNAs was determined by RT–PCR (ribosomal RNA as control) and western

blot analyses (tubulin as control). **C**, Binding of VG5Q to endothelial cells by cell adhesion assays. **D**, The t(5;11)(q13.3;p15.1) translocation associated with KTS increases expression of VG5Q. Results represent mean of triplicate cultures $\pm$ standard deviation from three independent experiments. EC, endothelial cell.

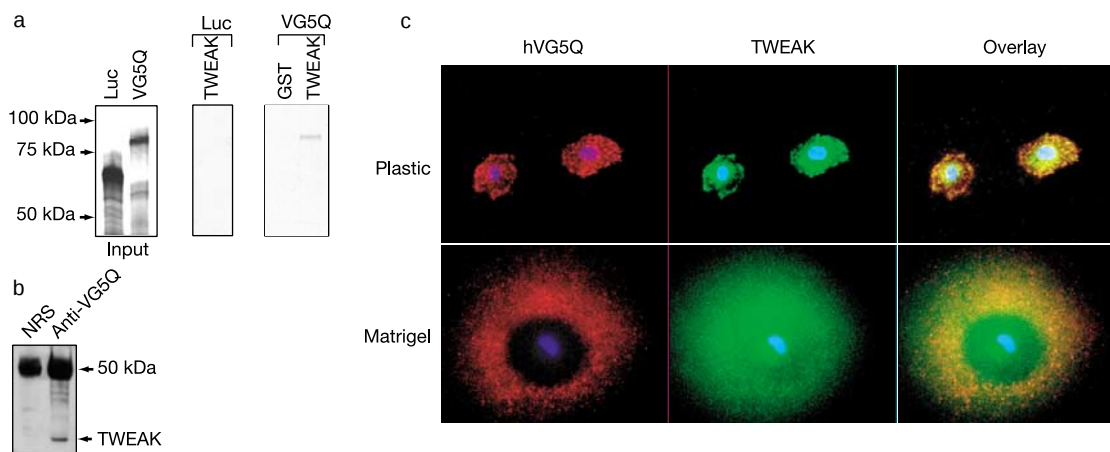


Figure 4 VG5Q interacts with TWEAK. **a**, Pull-down assays using GST-TWEAK. Input, ^{35}S -labelled luciferase (Luc, about 61 kDa) and VG5Q (about 84 kDa). Right panel, binding of VG5Q with GST-TWEAK, but not with GST; middle panel, no interaction between GST-TWEAK with luciferase. **b**, Co-immunoprecipitation of TWEAK with VG5Q from HSMC

protein extract by a rabbit anti-VG5Q antibody. Bound materials were analysed by western blot using a goat anti-TWEAK antibody. NRS, normal rabbit serum as a negative control. 50-kDa band, IgG cross-reaction. **c**, Co-localization of VG5Q and TWEAK in HUVECs. VG5Q, red signal; TWEAK, green; nuclei, blue (DAPI).

14 (Fn14), as a homotrimer, and it promotes angiogenesis *in vivo*²¹ as potently as VEGF and FGF-2, two well-known angiogenic factors^{1,18}. TWEAK treatment has been shown to promote cell proliferation and migration of HUVECs⁵. We demonstrated that VG5Q also induces proliferation of HUVECs (thymidine uptake: 680 ± 29 for wild-type VG5Q compared with 524 ± 14 for control (no VG5Q); $P = 0.007$, $n = 8$). The proliferation of HUVECs was also observed for mutant VG5Q with mutation E133K (thymidine uptake: 711 ± 37 compared with control (524 ± 14); $P = 0.001$, $n = 8$), but was not significantly different from wild-type VG5Q ($P > 0.05$, $n = 8$). As angiogenesis is a complex process involving endothelial cell protease secretion, proliferation, migration, adhesion and survival, a major effect of mutation E133K may be more prominent in processes other than proliferation.

KTS is genetically heterogeneous, and other KTS genes remain to be identified. We reported one *de novo* translocation, t(8;14)(q22.3;q13), associated with KTS²², and our mutation screening failed to identify a VG5Q mutation in the patient with this translocation. Molecular characterization of the t(8;14) translocation may lead to the identification of a new susceptibility gene for KTS. KTS is also expected to be phenotypically heterogeneous. Some KTS patients exhibit an absence of venous valves, hypoplasia of the veins, and complete absence of the deep venous system. Mutations in other KTS genes may account for such phenotypic variations.

Previously, a hypothesis of paradominant inheritance (autosomal lethal genes surviving only in a mosaic state) was proposed to explain the genetic basis of KTS and other syndromes that are characterized by sporadic occurrence and a mosaic distribution of lesions²³. On the basis of this model, patients with the VG5Q E133K mutation may carry a second mutational hit in VG5Q or another gene within the affected tissues. Alternatively, the sporadic nature of KTS may be explained by a model of autosomal-dominant inheritance with incomplete penetrance. Further studies are needed to distinguish between these hypotheses in the future when affected KTS tissues are available for molecular genetic analysis. Future studies will also determine whether VG5Q is involved in other pathological conditions associated with abnormal angiogenesis (tumour, ischaemia, inflammation).

As purified VG5Q promotes angiogenesis and suppression of VG5Q expression inhibits vessel formation, this work might have potentially important clinical therapeutic implications. VG5Q may be a target for angiogenic therapies (stimulating new blood vessel growth) or anti-angiogenic therapies (halting new blood vessel

growth) in treatment of angiogenesis-dependent disorders such as ischaemic heart disease and cancer. □

Methods

Construction of somatic cell hybrids

We constructed and analysed somatic cell hybrids derived from the blood sample of the t(5;11)(q13.3;p15.1) translocation patient as previously described²⁴.

Mutation analysis

Informed consent was obtained from the participants in accordance with the standards established by local institutional review boards. Genomic DNA was prepared from whole blood with the DNA Isolation Kit for Mammalian Blood (Roche Diagnostic). Single-strand conformation polymorphisms and DNA sequence analyses were carried out as described previously^{25–27}.

Western blot analysis

A polyclonal antibody against human/mouse VG5Q was developed using a synthetic polypeptide (LAQLRRKVEKLERELRSC) as the immunogen (Biosource International). The immunogen sequence corresponds to a unique portion of the amino terminus of VG5Q and did not match any other sequences in the databases, suggesting specificity of the VG5Q antibody. Western blot analysis was performed as described^{28,29}.

Expression and purification of human VG5Q protein

The full-length wild-type VG5Q cDNA was cloned into a bacterial expression vector pET-28b (Novagen), resulting in expression construct pET-28VG5Q-wt for 6 × His-tagged VG5Q. The VG5Q mutation E133K was introduced into pET-28VG5Q-wt using PCR-based site-directed mutagenesis, resulting in pET-28VG5Q-mt. The expression constructs were transformed into *Escherichia coli* BL21(DE3) Star, and 6 × His-VG5Q protein was purified using a Ni-NTA agarose column according to the manufacturer's instructions (Qiagen). The eluted protein was dialysed, and quality of purification was examined by SDS-polyacrylamide gel electrophoresis (PAGE) and western blot analysis.

Chick CAM assay

Fertilized chicken eggs were incubated at 37 °C for 4 days and then opened. The embryos were transferred into Petri dishes (100 mm diameter) and cultured at 37 °C for an additional 4 days to allow the development of CAM. Then, round-glass cellulose discs (3 mm diameter) soaked with either human recombinant VEGF (100 ng μl^{-1} ; Sigma, catalogue number V 7259) or with different concentrations of purified wild-type or mutant VG5Q (37.5 ng μl^{-1} , 75 ng μl^{-1} and 150 ng μl^{-1}) were placed on the CAMs. The control discs were soaked with the buffer that was used for dialysis and dissolving of VG5Q protein (50 mM Tris-HCl, 150 mM NaCl and 2 mM MgCl_2 , pH 7.4) or BSA (100 ng μl^{-1}). On the fifth day after the attachment of discs, the newly formed vessels were examined and visualized with a photomicroscope (Leica MZFLIII) and Spot Advanced software (Diagnostic Instruments). The number of newly formed vessels in the CAM was quantified with ImagePro Plus software (Media Cybernetics).

Competitive ELISA analysis

HUVECs were plated on Lab-Tek II chamber slides coated with or without matrigel (*in vitro* angiogenesis) for 4 h. The media were collected, incubated for 30 min with an optimum concentration (200 ng ml^{-1}) of anti-VG5Q antibody (determined experimentally with antigen), and transferred to wells coated with peptide immunogen (1 $\mu\text{g} \text{ ml}^{-1}$, 6 replicates). The bound antibody was detected by secondary horseradish

peroxidase (HRP)-conjugated donkey anti-rabbit IgG, chromogenic reaction, and absorbance reading of the wells. The negative controls include media only and PBS buffer. Positive control is the purified VG5Q protein. Blank media (media from matrigel-coated plates without cells) did not significantly reduce the HRP absorbance ($P = 0.24$) (Fig. 2G), indicating that decreases in absorbance with media from cells from matrigel culture are specific to VG5Q and are not due to the release of inhibitors from matrigel. The purified VG5Q did not give better inhibition in the ELISA assay because anti-VG5Q antibody was saturated.

Metabolic labelling

HUVECs were transiently transfected using HUVEC Nucleofector Kit (Amaxa Biosystems) with pcDNA3.1-VG5Q or vector pcDNA3.1, and metabolic labelling was carried out using Tran³⁵S-label (ICN Biochemicals) as described³⁰, but with modifications (see Supplementary Methods). Media from ³⁵S-labelled HUVEC transfectants were precipitated with an anti-VG5Q antibody, and precipitates were analysed by SDS-PAGE.

RNA interference

siRNA was selected 75 bases downstream from the start codon. The selected sequences were searched against the NCBI database to ensure that they were unique to VG5Q. The sequences for the two selected siRNAs are: 5'-AAUUGUCAUAGAUACACUGU-3' (siRNA1) and 5'-AAGAACAAAAAACUGGGAC-3' (siRNA2). The scrambled siRNA sequence is 5'-GCGCGCUUUGUAGGAUUCG-3'. The siRNA was synthesized by Dharmacon Research.

siRNA (1.6 nmol) was introduced into endothelial cells by transfection with oligofectamine (Invitrogen). Forty-eight hours after transfection, the adherent cells were collected by trypsinization. The cells were plated at a density of 0.2 million cells cm⁻² on Lab-Tek II chamber slides (Nalge Nunc International) coated with matrigel. Endothelial cell tube formation was examined 24 h later.

Cell adhesion assays

HUVECs were labelled with calcein AM (Molecular Probes), and incubated with 96-well plates coated with 2 µg ml⁻¹ of purified VG5Q or BSA (control). The unbound cells were removed by aspiration. Wells were washed with PBS, and read in a CytoFluor II Fluorescence Reader to measure fluorescence of adhering cells (see Supplementary Methods).

Protein-protein interactions

The cDNA segment encoding the extracellular domain of TWEAK (amino acids 93–249) was subcloned into pGEX-4T-1 to create the GST-TWEAK fusion protein. *In vitro* protein-protein interactions were carried out with pull-down assays using the GST-TWEAK fusion protein and *in-vitro*-translated, ³⁵S-labelled VG5Q produced in rabbit reticulocyte lysates using the TNT-coupled reticulocyte system (Promega). Co-immunoprecipitation was performed as described²⁸.

Cell proliferation assays

Endothelial cell proliferation assays were carried out with purified VG5Q using the [³H]thymidine-uptake method (see Supplementary Methods).

Received 26 September; accepted 23 December 2003; doi:10.1038/nature02320.

1. Yancopoulos, G. D. *et al.* Vascular-specific growth factors and blood vessel formation. *Nature* **407**, 242–248 (2000).
2. Berry, S. A. *et al.* Klippel-Trenaunay syndrome. *Am. J. Med. Genet.* **79**, 319–326 (1998).
3. Jacob, A. G. *et al.* Klippel-Trenaunay syndrome: spectrum and management. *Mayo Clin. Proc.* **73**, 28–36 (1998).
4. Wiley, S. R. *et al.* A novel TNF receptor family member binds TWEAK and is implicated in angiogenesis. *Immunity* **15**, 837–846 (2001).
5. Lynch, C. N. *et al.* TWEAK induces angiogenesis and proliferation of endothelial cells. *J. Biol. Chem.* **274**, 8455–8459 (1999).
6. Buschmann, I. & Schaper, W. Arteriogenesis versus angiogenesis: Two mechanisms of vessel growth. *News Physiol. Sci.* **14**, 121–125 (1999).
7. Carmeliet, P. & Jain, R. K. Angiogenesis in cancer and other diseases. *Nature* **407**, 249–257 (2000).
8. Folkman, J. & D'Amore, P. A. Blood vessel formation: what is its molecular basis? *Cell* **87**, 1153–1155 (1996).
9. Risau, W. Mechanisms of angiogenesis. *Nature* **386**, 671–674 (1997).
10. Whelan, A. J., Watson, M. S., Porter, F. D. & Steiner, R. D. Klippel-Trenaunay-Weber syndrome associated with a 5:11 balanced translocation. *Am. J. Med. Genet.* **59**, 492–494 (1995).
11. Durocher, D., Henckel, J., Fersht, A. R. & Jackson, S. P. The FHA domain is a modular phosphopeptide recognition motif. *Mol. Cell* **4**, 387–394 (1999).
12. Guglielmi, B. & Werner, M. The yeast homolog of human PinX1 is involved in rRNA and small nucleolar RNA maturation, not in telomere elongation inhibition. *J. Biol. Chem.* **277**, 35712–35719 (2002).
13. Hoggart, C. J. *et al.* Control of confounding of genetic associations in stratified populations. *Am. J. Hum. Genet.* **72**, 1492–1504 (2003).
14. Devlin, B., Roeder, K. & Wasserman, L. Genomic control, a new approach to genetic-based association studies. *Theor. Popul. Biol.* **60**, 155–166 (2001).
15. Shriver, M. D. *et al.* Ethnic-affiliation estimation by use of population-specific DNA markers. *Am. J. Hum. Genet.* **60**, 957–964 (1997).
16. Barker, D., Schafer, M. & White, R. Restriction sites containing CpG show a higher frequency of polymorphism in human DNA. *Cell* **36**, 131–138 (1984).
17. Pfeffer, S. Membrane domains in the secretory and endocytic pathways. *Cell* **112**, 507–517 (2003).
18. Mignatti, P., Morimoto, T. & Rifkin, D. B. Basic fibroblast growth factor, a protein devoid of secretory signal sequence, is released by cells via a pathway independent of the endoplasmic reticulum-Golgi complex. *J. Cell. Physiol.* **151**, 81–93 (1992).

19. Calvert, J. T. *et al.* Allelic and locus heterogeneity in inherited venous malformations. *Hum. Mol. Genet.* **8**, 1279–1289 (1999).
20. Baskerville, P. A., Ackroyd, J. S. & Browse, N. L. The etiology of the Klippel-Trenaunay syndrome. *Ann. Surg.* **202**, 624–627 (1985).
21. Wiley, S. R. & Winkles, J. A. TWEAK, a member of the TNF superfamily, is a multifunctional cytokine that binds the TweakR/Fn14 receptor. *Cytokine Growth Factor Rev.* **14**, 241–249 (2003).
22. Wang, Q. *et al.* Identification and molecular characterization of *de novo* translocation t(8;14)(q22.3;q13) associated with a vascular and tissue overgrowth syndrome. *Cytogenet. Cell Genet.* **95**, 183–188 (2001).
23. Happle, R. Klippel-Trenaunay syndrome: is it a paradigm trait? *Br. J. Dermatol.* **128**, 465–466 (1993).
24. Jackson, C. L., *et al.* in *Current Protocols in Human Genetics* (ed. Dracopoli, N. C.) 3.2.1–3.2.29 (Wiley, New York, 1996).
25. Chen, Q. *et al.* Genetic basis and molecular mechanism for idiopathic ventricular fibrillation. *Nature* **392**, 293–296 (1998).
26. Wang, Q. *et al.* SCN5A mutations associated with an inherited cardiac arrhythmia, long QT syndrome. *Cell* **80**, 805–811 (1995).
27. Wang, L., Fan, C., Topol, S. E., Topol, E. J. & Wang, Q. Mutation of MEF2A in an inherited disorder with features of coronary artery disease. *Science* **302**, 1578–1581 (2003).
28. Fan, C., Liu, M. & Wang, Q. Functional analysis of TBX5 missense mutations associated with Holt-Oram syndrome. *J. Biol. Chem.* **278**, 8780–8785 (2003).
29. Fan, C. *et al.* Novel TBX5 mutations and molecular mechanism for Holt-Oram syndrome. *J. Med. Genet.* **40**, e29 (2003).
30. Takagi, J., Petre, B. M., Walz, T. & Springer, T. A. Global conformational rearrangements in integrin extracellular domains in outside-in and inside-out signaling. *Cell* **110**, 599–611 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank E. Plow for critical reading of the manuscript; H. Yagita for plasmid hTWEAK/pCR3.1; S. Chen, M. K. Cathcart, A. Sedgehour, Z. Tang, C. Fan, P. B. Imrey, S. Archacki, D. Kitka, J. Poruban and A. Moore for help; and KTS patients for their participation in this study. This work was supported by NIH grants to Q.W.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to Q.W. (wangq2@ccf.org). The GenBank accession numbers are AY500994 for human VG5Q (hVG5Q) mRNA and amino acid sequences; AY500995 for mouse VG5Q (mVG5Q) mRNA and amino acid sequences; and AY500996 for human VG5Q genomic DNA sequence.

A conserved siRNA-degrading RNase negatively regulates RNA interference in *C. elegans*

Scott Kennedy, Duo Wang & Gary Ruvkun

Department of Molecular Biology, Massachusetts General Hospital and Department of Genetics, Harvard Medical School, Boston, Massachusetts 02114, USA

In many organisms, introducing double-stranded RNA (dsRNA) causes the degradation of messenger RNA that is homologous to the trigger dsRNA—a process known as RNA interference. The dsRNA is cleaved into short interfering RNAs (siRNAs), which hybridize to homologous mRNAs and induce their degradation¹. dsRNAs vary in their ability to trigger RNA interference: many mRNA-targeting dsRNAs show weak phenotypes, and nearly all mRNAs of the *Caenorhabditis elegans* nervous system are refractory to RNA interference^{2–4}. *C. elegans eri-1* was identified in a genetic screen for mutants with enhanced sensitivity to dsRNAs. Here we show that *eri-1* encodes an evolutionarily conserved protein with domains homologous to nucleic-acid-binding and exonuclease proteins. After exposure to dsRNA or siRNAs, animals with *eri-1* mutations accumulate more siRNAs than do wild-type animals. *C. elegans* ERI-1 and its human orthologue degrade siRNAs *in vitro*. In the nematode worm, ERI-1 is predominantly cytoplasmic and is expressed most highly in the

gonad and a subset of neurons, suggesting that ERI-1 siRNA activity suppresses RNA interference more intensely in these tissues. Thus, ERI-1 is a negative regulator that may normally function to limit the duration, cell-type specificity or endogenous functions of RNA interference.

We took advantage of the relative inefficiency of neuronal RNA interference (RNAi) in *C. elegans* to carry out a genetic screen for mutants with enhanced sensitivity in the nervous system to dsRNAs. Under normal conditions, such genes might be expected to inhibit the uptake or processing of dsRNAs, or to inhibit the amplification, spreading or stability of siRNAs. *unc-47*, encoding a probable GABA (γ -aminobutyric acid) transporter, is expressed in the 26 GABA-containing neurons of *C. elegans*⁵. Worms carrying an integrated *unc-47::GFP* fusion gene showed little or no decline in green fluorescent protein (GFP) fluorescence after feeding on bacteria expressing GFP dsRNA (Fig. 1).

We screened roughly 50,000 haploid genomes after ethyl methanesulphonate mutagenesis for mutants showing a marked decrease in the number of neurons that expressed GFP after feeding on *Escherichia coli* expressing GFP dsRNA. As a secondary screen, candidate mutants were tested for increased sensitivity to dsRNAs derived from endogenous chromosomal loci (see below). Of the 19 candidate mutants isolated from this genetic screen, two of the

strongest enhancers of RNAi defined the gene *eri-1* (named after enhanced RNAi). The Eri phenotypes of the two *eri-1* alleles, which are predicted to be null alleles, were indistinguishable, and both alleles showed low-frequency X-chromosome non-disjunction (a Him phenotype) and a highly penetrant temperature-sensitive sterile phenotype.

eri-1(mg366) mutants showed a wild-type pattern and intensity of *unc-47::GFP* fluorescence under normal growth conditions (see Supplementary Information for complete strain list). Unlike wild-type animals, *eri-1(mg366)* mutants feeding on bacteria expressing GFP dsRNA showed a 70% decrease in the number of GABAergic neurons exhibiting GFP fluorescence (Fig. 1a and Table 1). Notably, we observed an all or nothing response in individual neurons: GFP was either completely quenched by GFP dsRNA or unaffected by it (Fig. 1). The enhanced RNAi phenotype of *eri-1(mg366)* was tested with a second GFP fusion gene, *tub-1::GFP*, that is expressed in the sensory neurons (H. Mak, M. Basson and C. Johnson, personal communication). *tub-1::GFP* was not silenced in wild-type animals feeding on *E. coli* expressing GFP dsRNA, but *eri-1(mg366); tub-1::GFP* worms showed a 75% decrease in the number of neurons exhibiting GFP fluorescence after exposure to GFP dsRNA (Supplementary Fig. S1).

eri-1(mg366) also sensitized animals to dsRNA targeted against endogenous chromosomal loci, some of which do not function specifically in neurons. For example, *lin-1* mutations cause a multi-vulva (Muv) phenotype⁶, but for unknown reasons *lin-1* is refractory to RNAi in wild-type animals; however, *eri-1(mg366)* animals feeding on *E. coli* expressing *lin-1* dsRNA showed a Muv phenotype (Table 1). As compared with wild-type animals, *eri-1(mg366)* animals also showed enhanced sensitivity to *dpy-13*, *daf-19*, *myo-2*, *hmr-1*, *unc-86* and *daf-2* dsRNAs (Table 1). Two of these genes, *unc-86* and *daf-19*, are expressed exclusively in neurons. *eri-1* mutations did not enhance RNAi sufficiently in the nervous system to cause the expected loss-of-function phenotypes after RNAi of *unc-13*, *unc-17*, *unc-25* and *unc-47*. The *lin-1*, *dpy-13*, *myo-2*, *unc-22*

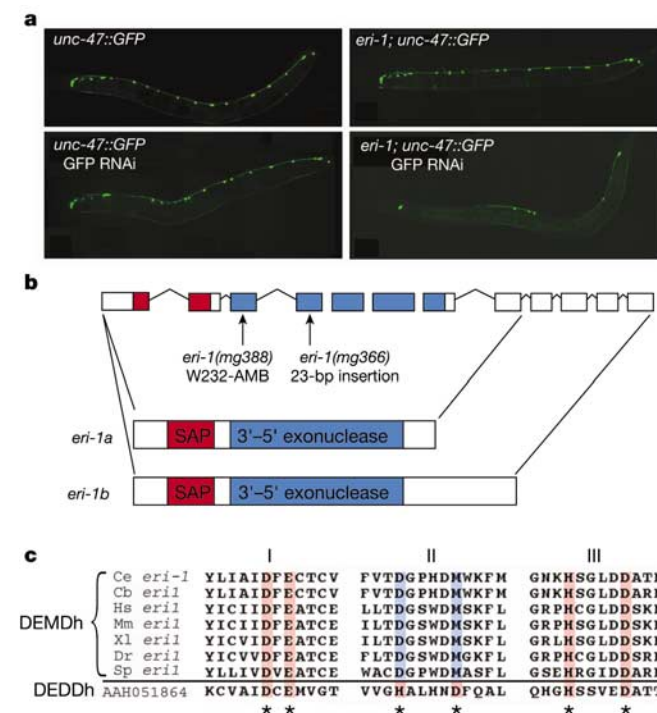


Figure 1 *eri-1* encodes a DEDDh exonuclease that enhances sensitivity to GFP dsRNA in GABAergic neurons. **a**, Wild-type and *eri-1(mg366)* L2 worms carrying an integrated *unc-47::GFP* transgene were grown on bacteria expressing either control vector (top) or dsRNA derived from GFP (bottom). Representative L4 progeny are shown. **b**, SAP/SAP-box and DEDDh 3'→5' exonuclease domains, shown in red and blue, respectively. The locations of the *mg366* and *mg388* lesions are indicated. **c**, Conserved motifs (I, II and III) found in the DEDDh superfamily of nucleases. Pink shading and asterisks indicate residues that are highly conserved in the DEDDh superfamily; blue shading indicates residues substituted in the *eri-1* DEDDh subfamily of nucleases. Cb, *Caenorhabditis briggsae* (AC084443); Hs, *Homo sapiens* (AAH35279); Mm, *Mus musculus* (NM_026067); Dr, *Danio rerio* (constructed by the fusion of BQ285328 and BI888174); Xl, *Xenopus laevis* (5' truncated EST; AW199662); Sp, *Schizosaccharomyces pombe* (NP_595533). AAH051864 is included for reference purposes.

Table 1 *eri-1* mutants show a general enhanced sensitivity to dsRNAs

Genotype	Fed dsRNA or injected siRNA*	Phenotype scored	% score
<i>unc-47::GFP</i>	Control vector	No. of GFP ⁺ neurons	19.8 ± 1.5
<i>eri-1(mg366); unc-47::GFP</i>	Control vector	No. of GFP ⁺ neurons	19.4 ± 1.5
<i>unc-47::GFP</i>	GFP dsRNA	No. of GFP ⁺ neurons	18.5 ± 2
<i>eri-1(mg366); unc-47::GFP</i>	GFP dsRNA	No. of GFP ⁺ neurons	6.6 ± 3
Wild type	<i>daf-19</i> dsRNA	% dauers	0
<i>eri-1(mg366)</i>	<i>daf-19</i> dsRNA	% dauers	12 ± 5
Wild type	<i>hmr-1</i> dsRNA	% lethality	13 ± 15
<i>eri-1(mg366)</i>	<i>hmr-1</i> dsRNA	% lethality	99 ± 2
Wild type	<i>daf-2</i> dsRNA	% dauers	0
<i>eri-1(mg366)</i>	<i>daf-2</i> dsRNA	% dauers	35 ± 5
Wild type	<i>unc-86</i> dsRNA	Uncoordinated	—
<i>eri-1(mg366)</i>	<i>unc-86</i> dsRNA	Uncoordinated	+
Wild type	<i>unc-22</i> siRNA*	Twitcher, day 1	16 ± 2
<i>eri-1(mg366)</i>	<i>unc-22</i> siRNA*	Twitcher, day 1	67 ± 7
Wild type	<i>unc-22</i> siRNA*	Twitcher, day 2	1 ± 1
<i>eri-1(mg366)</i>	<i>unc-22</i> siRNA*	Twitcher, day 2	55 ± 12
Wild type	<i>lin-1</i> dsRNA	% Muv	2 ± 0.2
<i>eri-1(mg366)</i>	<i>lin-1</i> dsRNA	% Muv	56 ± 2.9
<i>eri-1(mg366); rde-1(ne300)</i>	<i>lin-1</i> dsRNA	% Muv	0
<i>eri-1(mg366); rde-4(ne299)</i>	<i>lin-1</i> dsRNA	% Muv	0
<i>eri-1(mg366); mut-16(ne322)</i>	<i>lin-1</i> dsRNA	% Muv	0
<i>eri-1(mg366); sid-1(qt2)</i>	<i>lin-1</i> dsRNA	% Muv	0
Wild type	<i>dpy-13</i> dsRNA	Dumpy	+
<i>eri-1(mg366)</i>	<i>dpy-13</i> dsRNA	Dumpy	++++
<i>rrf-1(pk1417)</i>	<i>dpy-13</i> dsRNA	Dumpy	—
<i>eri-1(mg366); rrf-1(pk1417)</i>	<i>dpy-13</i> dsRNA	Dumpy	+

Worms of the indicated genotypes were grown on bacteria expressing the indicated dsRNA. In GFP RNAi experiments, L4 progeny were scored for GFP fluorescence in ventral cord neurons. All other dsRNA-expressing bacteria were obtained from the Ahringer library². For *dpy-13* RNAi, worms were scored as no dumpy (—), weak dumpy (+) or strong dumpy (++++). For *unc-86* RNAi, worms were scored as no *Unc* (—) and strong *Unc* (+). Results are the mean ± s.e.m. **unc-22* siRNAs (23 bp with 2 nt 3' overhangs) were microinjected and progeny from sequential egg lays (days 1–2) were scored¹⁵. Wild-type worms are N2 Bristol.

and *hmr-1* mRNAs are unlikely to be expressed in neurons; therefore, although *eri-1* was isolated in a screen for enhanced neuronal RNAi, loss of *eri-1* activity causes a generalized increase in the efficacy of RNAi in most tissues.

Genetic mapping localized *eri-1* to a one map unit region on the left arm of chromosome IV (Methods). Within this interval we identified a 23-base-pair (bp) insertion in the open reading frame T07A9.5 in *eri-1(mg366)*. This insertion encodes seven amino acids followed by a premature stop codon (Fig. 1b). A second independently isolated allele, *eri-1(mg388)*, has a point mutation in T07A9.5 that specifies a stop codon in place of Trp 231 (Fig. 1b). Both *eri-1(mg366)* and *eri-1(mg388)* are predicted to stop translation upstream of conserved domains (see below) and thus are likely to

show the null phenotype: a viable but temperature-sensitive sterile strain with enhanced RNAi. Transformation of wild-type T07A9.5 DNA into the *eri-1(mg366)* mutant rescued the sterility and enhanced RNAi phenotypes caused by *eri-1(mg366)* (Methods). Thus, T07A9.5 corresponds to *eri-1*.

Northern blot analysis indicated that *eri-1* encodes two equally abundant splice variants of roughly 1,400 and 1,800 nucleotides (nt), called *eri-1a* and *eri-1b*, respectively (Supplementary Fig. S2). Both splice variants encode a protein bearing a DEDDh-like 3'→5' exonuclease domain and a SAP/SAF-box domain (Fig. 1b). Members of the DEDDh family of nucleases include RNase T, oligoribonuclease and the proofreading subunit of *E. coli* DNA polymerase III (ref. 7). The DEDDh family member RNase T recognizes dsRNAs with 3' overhangs as preferred substrates⁸. *E. coli* oligoribonuclease (Orn) enzymes are required for the end degradation of mRNAs; *orn* mutants accumulate small mRNA degradation products of 2–5 nucleotides (nt) (ref. 9). The probable human orthologue of *eri-1* (called 3'hExo) has been biochemically purified as a factor that binds and degrades *in vitro* a 5-nt 3' overhang in the RNA stem-loop structure at the 3' end of a histone mRNA¹⁰. SAP/SAF-box domains show structural similarities to homeodomain DNA-binding proteins, and one member of this family can bind DNA¹¹. Although the SAP domain of ERI-1 suggests that ERI-1 has a possible function in the nucleus, the cytoplasmic localization of ERI-1 (see below) suggests that if ERI-1 has a nuclear function, this function is transient. It is also possible that, in the context of ERI-1, the SAP domain binds dsRNA to stabilize interactions between RNA and the nuclease domain.

Database searches identified a single probable ERI-1 orthologue in several vertebrate species and fission yeast that encode a SAP domain immediately amino-terminal to the exonuclease domain. Phylogenetic analysis using either the exonuclease domain or the SAP domain verified that these genes are orthologues (Supplementary Fig. S3). All of the *eri-1* orthologues contain a unique active site (Fig. 1c), including two substituted residues thought to be directly involved in catalysis in related exonucleases¹². Because the substituted residues are located in the active site, near where the RNA phosphodiester bond is broken by the nuclease, it is possible that ERI-1 and its orthologues have nucleic acid substrates that are related to those of RNase T and oligoribonuclease enzymes, but have

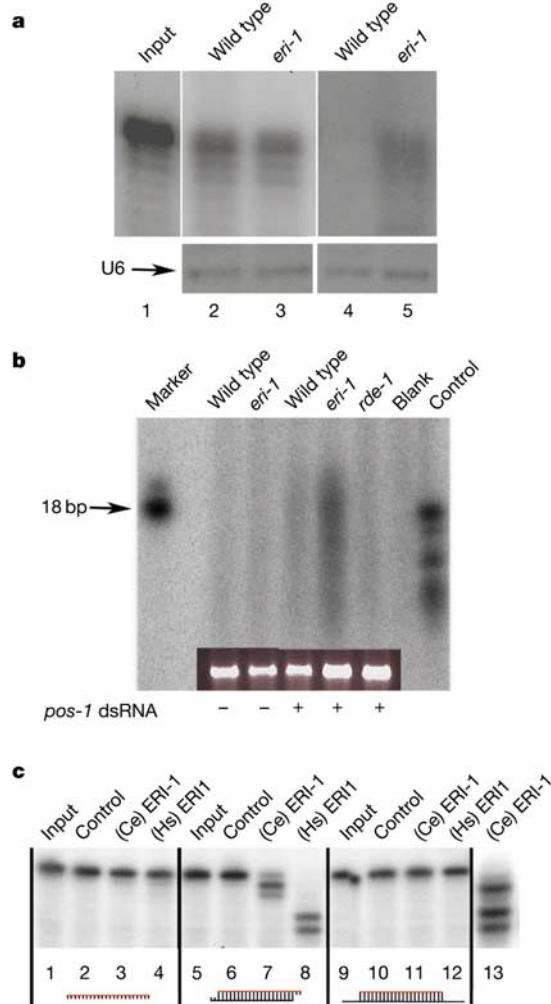


Figure 2 ERI-1 is an siRNase. **a**, Wild-type and *eri-1(mg366)* worms were injected with ³²P 5'-end-labelled *unc-22* double-stranded siRNAs (lane 1), and RNA was prepared from injected P0 worms (lanes 2 and 3) and their progeny (lanes 4 and 5). Blot was probed for U6 RNA as a loading control. **b**, RNA prepared from wild-type, *eri-1(mg366)* and *rde-1(ne300)* worms, grown for 14 h on control bacteria or bacteria expressing *pos-1* dsRNA diluted tenfold with control bacteria, was incubated with a *pos-1*-coding strand RNA probe and subjected to RNase protection. Also shown are 1 pmol of a 23-bp synthetic antisense *pos-1* siRNA (control) and an rRNA band from an agarose gel run in parallel. **c**, *In vitro* translated and immunoprecipitated *C. elegans* (Ce) ERI-1 and human (Hs) ERI1 or mock translation (control) were incubated as described¹⁰ with single-stranded *unc-22* siRNA (lanes 1–4), double-stranded *unc-22* siRNA containing 2-nt 3' overhangs (lanes 5–8 and 13) or *unc-22* siRNA hybridized to a 220-nt *unc-22* antisense RNA (lanes 9–12) for 30 min at either 37 °C (lanes 1–12) or 22 °C (lane 13). 5'-End-labelled siRNAs are shown in red.

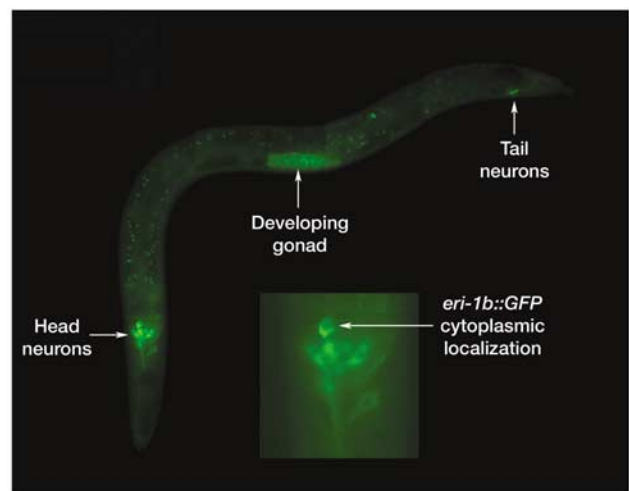


Figure 3 ERI-1 is localized in the cytoplasm of a subset of head and tail neurons. Dorsal is up and anterior is to the left. Fluorescence from the expression of a full-length *eri-1b::GFP* gene can be seen in a subset of head and tail neurons (including axonal projections) and in the developing gonad. Inset, magnification of GFP-expressing head neurons. A late L2 worm is shown.

distinctive features. We refer to this subfamily of DEDDh exonucleases as the 'DEMDh nucleases'.

In *Drosophila*, the RNase III enzyme Dicer cleaves dsRNAs into 21–23-nt double-stranded siRNAs with 2–4-nt 3' overhangs¹³. The 3' overhangs of siRNAs are required for efficient siRNA-mediated mRNA degradation¹⁴. Injection of a synthetic 23-bp *unc-22* siRNA with 2-nt 3' overhangs caused up to 55 times more progeny of the injected *eri-1(mg366)* mutant (–) animals than progeny of injected wild-type (*eri-1(+)*) animals to show an *unc-22* loss-of-function phenotype (Table 1), indicating that ERI-1 functions downstream of the processing of trigger dsRNA into siRNAs¹⁵. In addition, the physiological response to injected siRNAs was prolonged in *eri-1(–)* animals as compared with *eri-1(+)* animals (Table 1). Injected *unc-22* siRNAs were more abundant in the progeny of *eri-1(–)* animals than in the progeny of *eri-1(+)* animals (Fig. 2a). Similarly, progeny of *eri-1(–)* animals feeding on bacteria expressing dsRNA from the *pos-1* gene accumulated more *pos-1* siRNAs than did the progeny of *eri-1(+)* animals (Fig. 2b).

In vitro, *C. elegans* ERI-1 and its probable human ERI-1 orthologue partially degraded siRNAs with 2-nt 3' overhangs, but did not degrade single-stranded siRNAs or a single-stranded siRNA hybridized internally to a 220-nt *unc-22* message (Fig. 2c). *C. elegans* ERI-1 and human ERI1 produced similar reaction products when the reactions were done at temperatures at which the respective proteins function in nature (Fig. 2c, lanes 8 and 13). The increased sensitivity to siRNAs and increased abundance and stability of primary siRNAs in *eri-1* mutant animals, as well as the biochemical activity of ERI-1, suggest that ERI-1 inhibits RNAi by degrading the 3' overhangs of siRNAs. siRNAs lacking 3' overhangs may be non-functional because they fail to enter the RNAi-induced silencing complex (RISC). Alternatively, the ERI-1 3'-resected siRNAs may be unstable *in vivo*: they are not observed in wild-type animals after the injection of siRNAs (Fig. 2a). *In vivo*, additional nucleases, or possibly ERI-1 in conjunction with an RNA helicase, may catalyse the complete degradation of siRNAs.

ERI-1 is unlikely to function as a general exonuclease because *eri-1*-null mutants are viable and show limited pleiotropic phenotypes: that is, a weak Him phenotype and a temperature-sensitive sterility due to a defect in sperm function (data not shown). The only other known enhancer of RNAi in *C. elegans*, the RNA-dependent RNA polymerase encoded by *rrf-3*, shows identical pleiotropic phenotypes, suggesting that even these limited pleiotropies are probably a consequence of a defect in the siRNase (or possibly microRNase) function of ERI-1 and not in the processing of unrelated cellular mRNAs by ERI-1 (ref. 16). In addition, these shared pleiotropic phenotypes suggest that ERI-1 and RRF-3 may function in the same genetic pathway to inhibit RNAi. The ERI-1 orthologues may be dedicated siRNases and/or may have an additional RNA processing function, as suggested by the finding that *in vitro* human ERI-1 can process a histone mRNA stem loop¹⁰.

To determine where in the RNAi pathway *eri-1* functions, genetic epistasis analysis was carried out between *eri-1* and mutants defective in normal RNAi responses. We constructed double-mutant combinations containing *eri-1* and five genes required for RNAi: the RNAi-defective argonaute paralogue encoded by *rde-1*, the dsRNA-binding protein encoded by *rde-4*, the RNA-dependent RNA polymerase encoded by *rrf-1*, the RNAi-defective mutator gene *mut-16*, and the systemic RNAi-defective mutant *sid-1* (refs 17–20). *rde-1*, *rde-4*, *mut-16* and *sid-1* mutations were epistatic to *eri-1* for sensitivity to all dsRNAs tested (Table 1). By contrast, *eri-1* enhanced sensitivity to *hmr-1* and *dpy-13* dsRNAs in an *rrf-1*-null mutant background, suggesting that amplification of secondary siRNAs in somatic tissues is not essential for *eri-1* to enhance RNAi (ref. 20 and Table 1). Thus, consistent with the biochemical activity of ERI-1, the increased RNAi sensitivity of *eri-1* worms probably depends on the production of siRNAs by the canonical RNAi pathway.

To determine the tissues and subcellular compartment in which ERI-1 functions, we generated a fusion gene between GFP and the full-length ERI-1b protein, called *eri-1b::GFP* (see Supplementary Information). *eri-1* mutants stably expressing *eri-1b::GFP* were rescued for both enhanced RNAi and temperature-sensitive sterility, indicating that this fusion gene was functional and representative of endogenous *eri-1* expression. In three independent lines, we observed GFP expression in the developing somatic gonad and in a subset of neurons (Fig. 3). In adults, ERI-1 was expressed in neurons and gonadal expression was restricted to the spermatheca (Supplementary Fig. S4). In neurons, ERI-1 is localized predominantly to the cytoplasm (Fig. 3). The expression and the cytoplasmic localization of the *eri-1::GFP* fusion was not markedly altered by exposure to dsRNA (data not shown).

Fusion of GFP to the predicted promoter of *eri-1* (termed *eri-1P::GFP*), which was expected to reveal the pattern of expression of both *eri-1a* and *eri-1b*, showed a pattern of expression similar to that of *eri-1b::GFP*. This fusion protein containing only the first five residues of ERI-1 was not localized preferentially to the cytoplasm. Using this construct, we also observed low ubiquitous expression throughout the worm (Supplementary Fig. S4). The high expression of ERI-1 in a subset of neurons may, at least in part, explain the relative inefficiency of RNAi in these neurons in wild-type worms. The expression of *eri-1* in the spermatheca may be related to the defect in sperm function observed in *eri-1* mutants. Interestingly, *spe* genes that specify steps in the maturation of *C. elegans* sperm are also resistant to RNAi (ref. 2). The low ubiquitous expression revealed by *eri-1P::GFP* may explain the observed generalized increase in the efficacy of RNAi observed in *eri-1* mutants.

The identification of *eri* loci indicates that the RNAi machinery is under substantial negative regulation. This negative regulation may limit the intensity of an episode of RNAi, inhibit RNAi in particular cell types, and/or regulate the endogenous functions of the RNAi machinery. Inhibition of the *eri-1* orthologues, for example by drugs that target their unique DEMDh active site, may facilitate more efficient RNAi in mammals and fungi. □

Methods

Cloning of *eri-1*

We mapped *eri-1(mg366)* using a Hawaiian isolate of *C. elegans* (CB4856). A total of 141 F₂ recombinants were scored for *eri-1* on the basis of sensitivity to GFP dsRNA, sensitivity to *hmr-1* dsRNA, and/or the temperature-sensitive sterility phenotype. Single-nucleotide polymorphism mapping established a right boundary on chromosome IV corresponding to a genetic map position of about –23.1. Because of a direct repeat, the exact sequence of the 23-bp insertion in *mg366* is unknown; however, it includes 23 of the following 32 nt: ttctgaatagtcctgttttttttcgataaa. *mg388* is a G to A transition at nucleotide position 35,375 of cosmid T07A9.

eri-1 rescue

T07A9.5 is downstream in an operon with *daf-18*. Transformation of DNA containing the predicted *daf-18* and T07A9.5 operon into *eri-1(mg366)* mutants rescued the temperature-sensitive sterility and enhanced RNAi phenotypes caused by *eri-1(mg366)*. Transformation of T07A9.5 alone, including 5' sequences up to the *daf-18* locus, did not rescue *eri-1(mg366)*, consistent with the notion that T07A9.5 is co-transcribed with *daf-18*. Transformation of T07A9.5 expressed under the control of the ubiquitously expressing heterologous *sur-5* promoter²¹ rescued the enhanced RNAi and temperature-sensitive sterility of *eri-1(mg366)*. Transgenic DNA was generated by pooling six independent polymerase chain reactions from N2 animals. T07A9.5 operon rescuing DNA corresponds to cosmid T07A9 positions 43,255 to 32,125. DNA was injected into *eri-1(mg366)* at 5 ng µl^{–1} with 20 ng µl^{–1} of *tub-1::GFP* marker DNA. Control lines generated with *tub-1::GFP* marker DNA were non-rescuing. A complete description of transgene construction is given in the Supplementary Information.

In vitro activity of *eri-1*

C. elegans eri-1A and human *eri-1* complementary DNAs were appended with a carboxy-terminal T7 promoter and an N-terminal Flag epitope and transcribed *in vitro*. mRNAs were translated in reticulocyte lysates and immunoprecipitated with agarose beads conjugated to anti-Flag antibodies and eluted with a Flag peptide.

Received 5 November; accepted 19 December 2003; doi:10.1038/nature02302.

1. Hannon, G. J. RNA interference. *Nature* **418**, 244–251 (2002).

2. Kamath, R. S. et al. Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* **421**, 231–237 (2003).

3. Timmons, L., Court, D. L. & Fire, A. Ingestion of bacterially expressed dsRNAs can produce specific and potent genetic interference in *Caenorhabditis elegans*. *Gene* **263**, 103–112 (2001).
4. Tavernarakis, N., Wang, S. L., Dorovkov, M., Ryazanov, A. & Driscoll, M. Heritable and inducible genetic interference by double-stranded RNA encoded by transgenes. *Nature Genet.* **24**, 180–183 (2000).
5. McIntire, S. L., Reimer, R. J., Schuske, K., Edwards, R. H. & Jorgensen, E. M. Identification and characterization of the vesicular GABA transporter. *Nature* **389**, 870–876 (1997).
6. Beitel, G. J., Tuck, S., Greenwald, I. & Horvitz, H. R. The *Caenorhabditis elegans* gene *lin-1* encodes an ETS-domain protein and defines a branch of the vulval induction pathway. *Genes Dev.* **9**, 3149–3162 (1995).
7. Zuo, Y. & Deutscher, M. P. Exoribonuclease superfamilies: structural analysis and phylogenetic distribution. *Nucleic Acids Res.* **29**, 1017–1026 (2001).
8. Li, Z., Pandit, S. & Deutscher, M. P. 3' exoribonucleolytic trimming is a common feature of the maturation of small, stable RNAs in *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **95**, 2856–2861 (1998).
9. Ghosh, S. & Deutscher, M. P. Oligoribonuclease is an essential component of the mRNA decay pathway. *Proc. Natl Acad. Sci. USA* **96**, 4372–4377 (1999).
10. Dominski, Z., Yang, X. C., Kaygun, H., Dadlez, M. & Marzluff, W. F. A 3' exonuclease that specifically interacts with the 3' end of histone mRNA. *Mol. Cell* **12**, 295–305 (2003).
11. Kipp, M. *et al.* SAF-box, a conserved protein domain that specifically recognizes scaffold attachment region DNA. *Mol. Cell Biol.* **20**, 7480–7489 (2000).
12. Hamdan, S., Carr, P. D., Brown, S. E., Ollis, D. L. & Dixon, N. E. Structural basis for proofreading during replication of the *Escherichia coli* chromosome. *Structure Fold. Des.* **10**, 535–546 (2002).
13. Zamore, P. D., Tuschl, T., Sharp, P. A. & Bartel, D. P. RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell* **101**, 25–33 (2000).
14. Elbashir, S. M., Martinez, J., Patkaniowska, A., Lendeckel, W. & Tuschl, T. Functional anatomy of siRNAs for mediating efficient RNAi in *Drosophila melanogaster* embryo lysate. *EMBO J.* **20**, 6877–6888 (2001).
15. Caplen, N. J., Parrish, S., Imani, F., Fire, A. & Morgan, R. A. Specific inhibition of gene expression by small double-stranded RNAs in invertebrate and vertebrate systems. *Proc. Natl Acad. Sci. USA* **98**, 9742–9747 (2001).
16. Simmer, F. *et al.* Loss of the putative RNA-directed RNA polymerase RRF-3 makes *C. elegans* hypersensitive to RNAi. *Curr. Biol.* **12**, 1317–1319 (2002).
17. Tabara, H. *et al.* The *rde-1* gene, RNA interference, and transposon silencing in *C. elegans*. *Cell* **99**, 123–132 (1999).
18. Tabara, H., Yigit, E., Siomi, H. & Mello, C. C. The dsRNA binding protein RDE-4 interacts with RDE-1, DCR-1, and a DExH-box helicase to direct RNAi in *C. elegans*. *Cell* **109**, 861–871 (2002).
19. Winston, W. M., Molodowitch, C. & Hunter, C. P. Systemic RNAi in *C. elegans* requires the putative transmembrane protein SID-1. *Science* **295**, 2456–2459 (2002).
20. Sijen, T. *et al.* On the role of RNA amplification in dsRNA-triggered gene silencing. *Cell* **107**, 465–476 (2001).
21. Gu, T., Orita, S. & Han, M. *Caenorhabditis elegans* SUR-5, a novel but conserved protein, negatively regulates LET-60 Ras activity during vulval induction. *Mol. Cell Biol.* **18**, 4556–4564 (1998).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank J. Roig, G. Lenz and J. Avruch for reagents and advice; A. Frand for critically evaluating our manuscript; and members of the Ruvkun laboratory for advice and discussions.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to G.R. (ruvkun@molbio.mgh.harvard.edu).

Cytoplasmic dynein functions as a gear in response to load

Roop Mallik¹, Brian C. Carter¹, Stephanie A. Lex², Stephen J. King² & Steven P. Gross¹

¹Department of Developmental and Cell Biology, University of California Irvine, Irvine, California 92612, USA

²Division of Molecular Biology and Biochemistry, School of Biological Sciences, University of Missouri-Kansas City, Kansas City, Missouri 64110, USA

Cytoskeletal molecular motors belonging to the kinesin and dynein families transport cargos (for example, messenger RNA, endosomes, virus) on polymerized linear structures called microtubules in the cell¹. These 'nanomachines' use energy obtained from ATP hydrolysis to generate force², and move in a step-like manner on microtubules. Dynein^{3–5} has a complex and funda-

mentally different structure from other motor families. Thus, understanding dynein's force generation can yield new insight into the architecture and function of nanomachines. Here, we use an optical trap⁶ to quantify motion of polystyrene beads driven along microtubules by single cytoplasmic dynein motors. Under no load, dynein moves predominantly with a mixture of 24-nm and 32-nm steps. When moving against load applied by an optical trap, dynein can decrease step size to 8 nm and produce force up to 1.1 pN. This correlation between step size and force production is consistent with a molecular gear mechanism. The ability to take smaller but more powerful strokes under load—that is, to shift gears—depends on the availability of ATP. We propose a model whereby the gear is downshifted through load-induced binding of ATP at secondary sites in the dynein head.

Cytoplasmic dynein is responsible for most intracellular retrograde (towards microtubule minus ends) transport as well as chromosome segregation during mitosis^{4,5}. To understand its function in a controlled setting, purified cytoplasmic dynein was adsorbed on beads that served as markers for motor position⁷ (see Methods). An optical trap was used to place dynein-coated beads on immobilized microtubules. The microtubule-binding probability of beads followed single-molecule Poisson statistics, suggesting that processive motion was driven by single dynein molecules⁸ (Fig. 1a; see also Supplementary Fig. 2). Motor function was quantified using video tracking⁹ and an optical trap with a quadrant photodiode detector¹⁰ (see Methods). The force exerted by the motor was determined from the product of trap stiffness and displacement

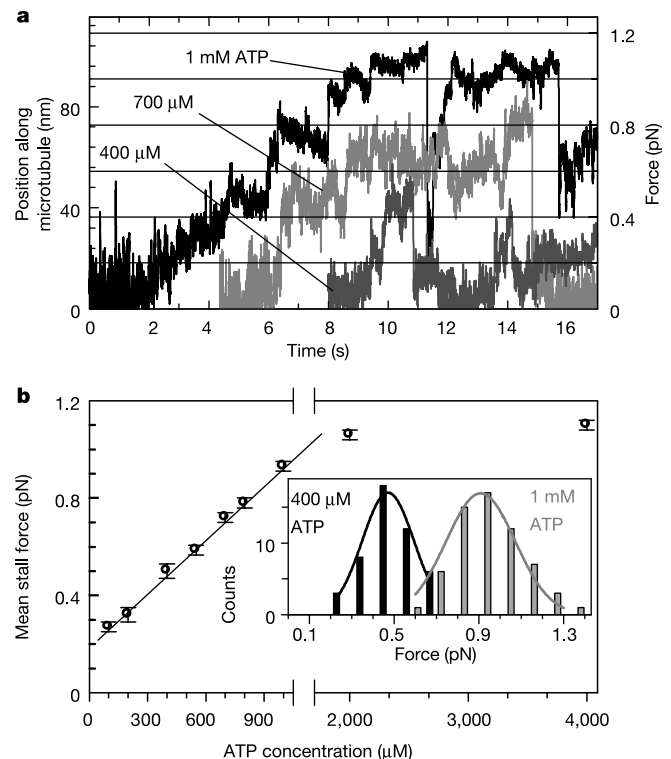


Figure 1 Stall force of cytoplasmic dynein. **a**, Quadrant photodetector record of dynein-driven bead motion in the optical trap, at different ATP concentrations. Time axis origin is adjusted for clarity. **b**, Mean stall force ($\pm$ s.e.m.) from multiple displacement records as a function of ATP. A straight line emphasizes linearity of the data below 1 mM ATP. Inset, distribution of stall force at 400 μ M and 1 mM ATP. Solid lines show gaussian fits to the respective data (400 μ M ATP, mean = 0.50 pN, σ = 0.26 pN, n = 47; 1 mM ATP, mean = 0.92 pN, σ = 0.31 pN, n = 62).

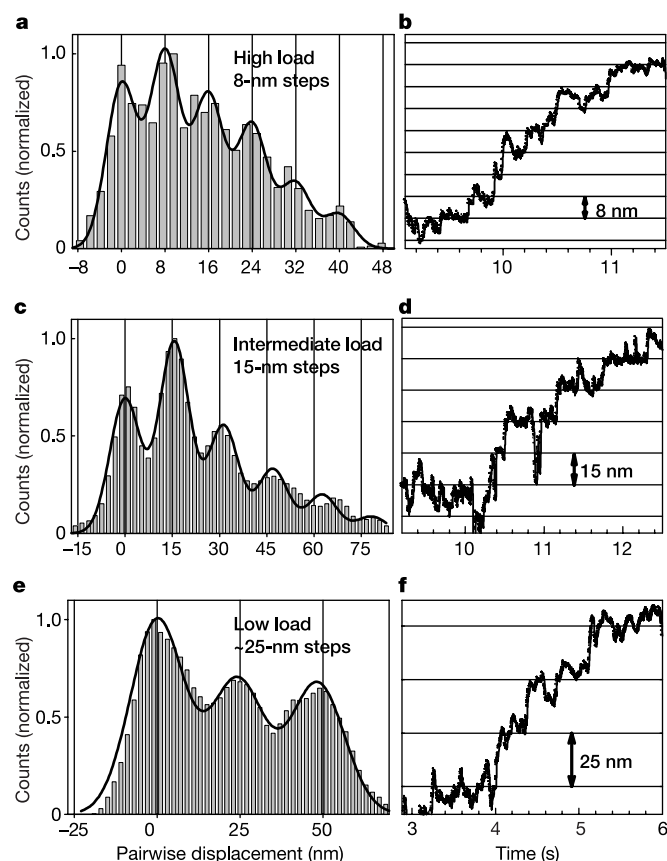


Figure 2 Dynein takes shorter steps under load. **a**, Pairwise distance function (PDF) of displacements under load approaching stall (>0.8 pN). The solid line is fitted to the sum of six gaussians, $i = -2$ to 3 , $\sum_i A_i \exp[-(x - i\Delta)^2/(2\sigma^2)]$, where Δ is the unit step size, A_i is the amplitude and σ is the standard deviation. After correction for bead-linkage stiffness (see Methods), $\Delta = 7.9$ nm, $\sigma = 2.8$ nm. **b**, Eight-nanometre steps in the displacement record. Data are median filtered with a 10-ms time window. Trap stiffness = 0.011 pN nm $^{-1}$. **c**, PDF of displacements at intermediate load (about 0.40 – 0.80 pN). The solid line is fitted to the sum of six gaussians, $\Delta = 15.5$ nm, $\sigma = 4.8$ nm. **d**, Fifteen-nanometre steps. Trap stiffness = 0.007 pN nm $^{-1}$. **e**, PDF of displacements at low load (<0.4 pN). The solid line is fit to the sum of three gaussians, $\Delta = 25.3$ nm, $\sigma = 8.1$ nm. **f**, Twenty-five-nanometre steps. Trap stiffness = 0.007 pN nm $^{-1}$. Each plot is a representative example of multiple recordings. In a given load range, a mixture of different step sizes cannot be ruled out.

of the bead from the trap centre, after correction for bead–microtubule linkage stiffness⁸ (see Methods).

Plotted as a function of ATP concentration (Fig. 1b), the mean stall force increases linearly up to 1 mM ATP, beyond which it reaches a maximum value of 1.1 pN. At 100 μ M ATP the maximum force produced is only about 0.25 pN, a reduction of approximately 75% from the stall force value at saturating (4 mM) ATP. The magnitude of this reduction is much larger than for kinesin¹¹. The distribution of stall forces at 400 μ M and 1 mM ATP (Fig. 1b, inset) demonstrates a population shift from ‘weaker’ to ‘stronger’ motor species as the ATP concentration increases. The maximal stall force of 1.1 pN at saturating ATP is similar to that observed for axonemal dynein from inner arm of *Chlamydomonas* flagella¹², and is consistent with the predicted *in vivo* stall force of 1.1 ± 0.04 pN for a single cytoplasmic dynein¹³.

To investigate further motor function at high ATP, the size of unitary steps of dynein was obtained from the pairwise distance function (PDF) of bead displacement records¹⁴. A representative

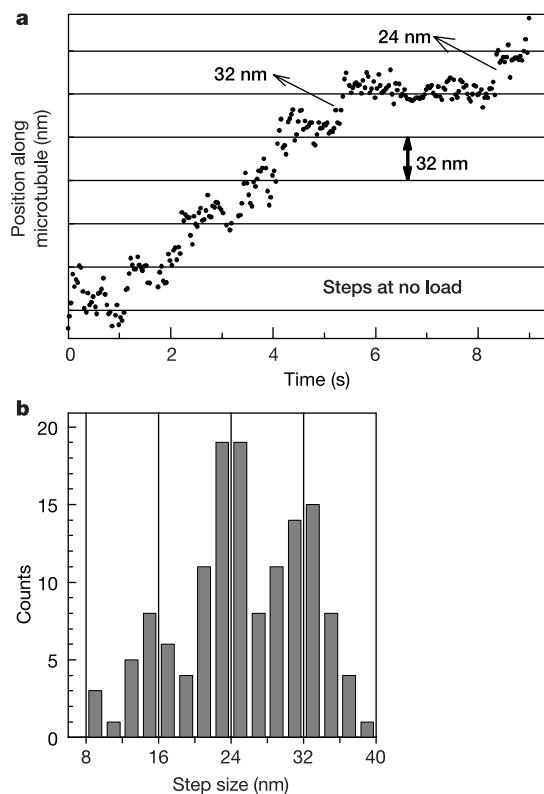


Figure 3 Cytoplasmic dynein takes predominantly 24- and 32-nm steps under no load. **a**, Representative video track of bead driven by single dynein molecule at 2 μ M ATP. The displacement parallel to the microtubule axis is plotted. One 24-nm and another 32-nm step are marked. **b**, Histogram showing the distribution of step sizes under no load. Vertical grids mark the microtubule lattice periodicity of 8 nm.

histogram of PDF plots from the high-load regime (load >0.8 pN) shows a periodicity of 8 nm in the displacement record (Fig. 2a). Consecutive 8 -nm steps in the raw displacement record are also shown (Fig. 2b). Thus, under high load cytoplasmic dynein takes steps of 8 nm. As dynein function might be regulated through variation in step size, we calculated the PDF for displacement under intermediate load conditions (about 0.4 to 0.8 pN). A periodicity of 15 nm was found (Fig. 2c, d). Such a step size was predicted for axonemal dynein from electron microscopic studies¹⁵. Under conditions approaching zero load (<0.4 pN), we observed a periodicity of approximately 25 nm (Fig. 2e, f). Note that this is the same bead displacement record that yielded 15 -nm steps at higher load (see above). Thus, the same motor steps in approximately 25 -nm increments when under low load, and then reduces step size to 15 nm as it encounters increasing load moving away from the trap centre. Because of the large (about 25 nm) step size, only a limited number of steps occur in the optical trap before the increasing load presumably effects a transition to smaller steps. To investigate zero-load function further, we used video tracking⁹ of unloaded beads (no optical trap). This tracking system was tested by moving dynein-bound beads with nanometre-sized, Poisson-distributed steps (see Supplementary Information).

At low (a few μ M) ATP concentrations, ATP binding becomes rate limiting and individual motor steps can be detected with the temporal resolution of video tracking^{9,16}. We observed a complex admixture of step sizes, with steps frequently close to 24 or 32 nm (Fig. 3a). The histogram of step sizes from multiple measurements (Fig. 3b; see also Supplementary Information) shows peaks at about

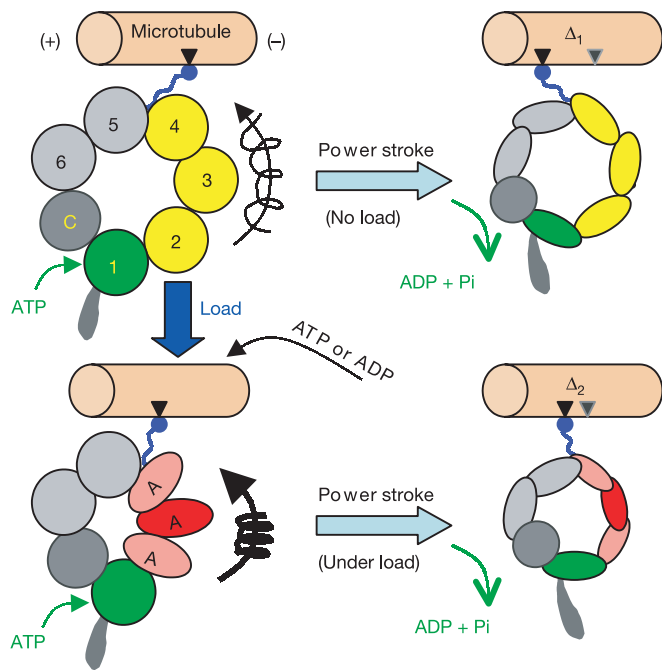


Figure 4 Model for an ATP-regulated gear. The dynein head is shown as six sequential AAA domains (numbered spheres) and one carboxy-terminal subdomain (C) connected in a ring conformation with a projecting microtubule-binding stalk (blue). P_i , phosphate group. Top panel: zero load. Primary ATP hydrolysis (at AAA number 1, green) causes propagating conformational change through domains 2–4 (yellow) leading to power stroke (left, before; right, after). The step size (Δ_1) is shown as distance between binding positions of the stalk on the microtubule (vertical arrowheads) before and after the power stroke. Bottom panel: function under load. Load-induced nucleotide (A indicates ATP or ADP) binding (at AAA number 3, dark red; possible also at numbers 2 and 4, light red) compacts ring conformation (note smaller diameter), resulting in a stiffer linkage (thick black arrow with spring) that can transmit higher force. The step size is now smaller ($\Delta_2 < \Delta_1$).

15 nm, 24 nm and 32 nm. The combination of 24- and 32-nm steps might result from attempted steps of an intermediate magnitude (possibly 30 nm; see Supplementary Information) modulated by the microtubule lattice. Weak electrostatic interactions with the microtubule-binding sites could impose a backward (24-nm step) or forward (32-nm step) diffusive bias on the dynein head. Although only 8-nm steps are reported for axonemal dynein, published studies have characterized function only under high load^{12,17}. Interestingly, axonemal dynein arms can produce oscillatory displacements¹⁸ of a singlet microtubule with an amplitude of about 30 nm, which might reflect single back-and-forth steps with 30-nm magnitude close to the unloaded step size reported here (see Supplementary Information).

The distribution of residence times at low ATP shows a single-exponential decay (see Supplementary Information), and thus favours a 1:1 coupling scenario (single ATP hydrolysis gives single step) under no externally applied load¹⁹. With our data it is not possible to determine the coupling ratio at high load. However, the observed decrease in step size under load may indicate that the coupling remains unchanged, with the reduced step size resulting in a larger force—in other words, a gear mechanism. This is radically different from force production in kinesin¹⁴ or myosin²⁰, where step size under load is invariant. For a motor protein, the work done per step is: $W_{\text{step}} = \text{step size} \times \text{stall force}$. Using a step size of 32 nm and stall force of about 0.3 pN in the low-load regime, W_{step} is about 9.6 pN nm. At 4 mM ATP, with 8-nm step and 1.1-pN stall, W_{step} is about 8.8 pN nm. As W_{step} is almost unchanged (within experi-

mental error), the increase in stall force can be effected by the observed reduction in step size. Thus, from an energy conservation point of view there is no requirement of multiple ATP hydrolysis per step under load. About 90 pN nm of free energy is available from hydrolysis of a single ATP, so assuming 1:1 coupling, cytoplasmic dynein is roughly 10% efficient, in contrast to the approximately 50% efficiency of kinesin⁸.

How could such a nanoscale gear be implemented? Dynein is an AAA protein^{4,21,22} (ATPase associated with diverse cellular activities), with complex architecture. It has four AAA domains (numbered 1 to 4), each of which can potentially bind ATP^{4,23} (Fig. 4; see also Supplementary Information). It has been suggested^{22,24,25} that nucleotide-binding-induced changes in relative orientation between adjacent AAA domains can lead to a 'tighter' conformation of the AAA ring (see Supplementary Information). Evidence for such ring compaction has been found in the NSF (*N*-ethylmaleimide-sensitive factor) AAA proteins²⁶. We propose a dynein gear based on such a mechanism (Fig. 4), where load-induced ATP binding tightens the AAA ring, and yields a shorter but stronger step of the motor. In this extension of previous models^{3,22}, ATP has the novel role of regulator of dynein step size, as well as being an energy input in the chemomechanical cycle. Could it be that dynein always generates a constant force of 1.1 pN, but under applied opposing load is unable to reach a distant binding site? Such a scenario does not agree with the stall force reduction at low ATP (Fig. 1b), where the motor is never observed to exert 1.1 pN force. The shortening of step size seems to arise out of a need to produce higher force, with the force-producing machinery (proposed gear mechanism) being disrupted by the unavailability of ATP. (See Supplementary Information for more information about the model.)

What are the implications of such a gear? Compared with the 8-nm steps in kinesin, at low load the 32-nm steps of dynein imply that as a cargo transporter, dynein is four times more fuel-efficient than kinesin. Furthermore, the ATP-dependent gear machinery of dynein is a potential target for dynein regulation. The single-molecule measurements presented here address a fundamental question in the field of molecular motors: why is the architecture of dynein so complex and different from that of kinesin or myosin? Our results suggest that this complexity implements a nanoscale transmission that enables dynein to shift gears, and minimize ATP consumption while providing force tailored to overcome external load. It will be an exciting challenge to apply the mechanical insights presented here to better understand other important AAA machines. □

Methods

Protein purification

Cytoplasmic dynein from bovine brain tissue was purified and isolated from dynactin as described²⁷. Aliquots were flash frozen and stored at -80°C . Before motility assay, dynein was further purified by an ATP-sensitive microtubule affinity assay²⁸. Bovine brain tubulin was purified over a phosphocellulose column²⁹. Absence of any significant contamination from dynactin was established through SDS-polyacrylamide gel electrophoresis (PAGE) gel (Supplementary Fig. 1).

In vitro motility assay

Assays were performed at 24°C in a flow chamber (4×25 mm, volume approximately 8 μl) made with cleaned, polylysine-coated coverslips. Taxol-stabilized microtubules were immobilized on the coverslip surface before casein-blocking of surface. Dynein was first incubated with carboxylated polystyrene beads (450-nm diameter; Polysciences Inc.) in assay buffer (ATP plus 0.3×35 mM PIPES, pH 7.2, 5 mM MgSO_4 , 1 mM EGTA, 0.5 mM EDTA, plus 1 mM GTP plus 20 μM Taxol). Further binding was blocked with casein (5 mg ml^{-1} in assay buffer). Beads were viewed with video-enhanced differential interference contrast microscopy in an inverted microscope (modified Nikon TE-200). Custom-built image-processing software (Labview 6.1, National Instruments) was used for video tracking. Subpixel resolution in video tracking was confirmed as described¹⁵. Dynein-coated beads were brought into contact with microtubules using the optical trap. The fraction of beads binding to and moving on microtubules, when scored as a function of dynein:bead molar ratio, obeyed single-molecule Poisson statistics^{8,12,14,17,30} (Supplementary Fig. 2). Assays were performed at a bead-to-microtubule binding fraction of 0.3 or less. At this concentration, taking into account geometric constraints applicable

to a random distribution⁸ of dynein molecules on the bead, >99% of beads are driven by single dynein molecules.

Optical trapping was done as described^{6,10} using an 830-nm single-mode diode laser (Melles Griot). Trap stiffness was calculated from the power spectrum of thermal fluctuations of trapped beads¹⁰. Back focal plane imaging of bead position in the optical trap was done with a quadrant photodiode detector¹⁰. The detector output was anti-alias filtered at 1 kHz and digitized at 2 kHz. The correction factor to the bead displacement^{8,12,17} (about 1.06) due to elasticity of the bead–dynein linkage was determined from the variance of beads bound by motor to the microtubule in the presence of ADP, and from the trap stiffness.

Received 8 September; accepted 11 December 2003; doi:10.1038/nature02293.

- Hirokawa, N. Kinesin and dynein superfamily proteins and the mechanism of organelle transport. *Science* **279**, 519–526 (1998).
- Hackney, D. D. The kinetic cycles of myosin, kinesin and dynein. *Annu. Rev. Physiol.* **58**, 731–750 (1996).
- Asai, D. J. & Koonce, M. P. The dynein heavy chain: structure, mechanics and evolution. *Trends Cell Biol.* **11**, 196–202 (2001).
- King, S. M. AAA domains and organization of the dynein motor unit. *J. Cell Sci.* **113**, 2521–2526 (2000).
- King, S. M. The dynein microtubule motor. *Biochim. Biophys. Acta* **1496**, 60–75 (2000).
- Rice, S. E. & Spudis, J. A. Building and using optical traps to study properties of molecular motors. *Methods Enzymol.* **361**, 112–133 (2003).
- Wang, Z., Khan, S. & Sheetz, M. P. Single cytoplasmic dynein molecule movements: characterization and comparison with kinesin. *Biophys. J.* **69**, 2011–2023 (1995).
- Svoboda, K. & Block, S. M. Force and velocity measured for single kinesin molecules. *Cell* **77**, 773–784 (1994).
- Gelles, J., Schnapp, B. J. & Sheetz, M. P. Tracking kinesin-driven movements with nanometre-scale precision. *Nature* **331**, 450–453 (1988).
- Visscher, K. & Block, S. M. Versatile optical traps with feedback control. *Methods Enzymol.* **298**, 460–489 (1998).
- Visscher, K., Schnitzer, M. J. & Block, S. M. Single kinesin molecules studied with a molecular force clamp. *Nature* **400**, 184–189 (1999).
- Sakakibara, H., Kojima, H., Sakai, Y., Katayama, E. & Ooiwa, K. Inner-arm dynein c of *Chlamydomonas* flagella is a single-headed processive motor. *Nature* **400**, 586–590 (1999).
- Gross, S. P., Welte, M. A., Block, S. M. & Wieschaus, E. F. Dynein-mediated cargo transport *in vivo*: A switch controls travel distance. *J. Cell Biol.* **148**, 945–955 (2000).
- Svoboda, K., Schmidt, C. F., Schnapp, B. J. & Block, S. M. Direct observation of kinesin stepping by optical trapping interferometry. *Nature* **365**, 721–727 (1993).
- Burgess, S. A., Walker, M. L., Sakakibara, H., Knight, P. J. & Ooiwa, K. Dynein structure and power stroke. *Nature* **421**, 715–718 (2003).
- Hua, W., Young, E. C., Fleming, M. L. & Gelles, J. Coupling of kinesin steps to ATP hydrolysis. *Nature* **388**, 390–393 (1997).
- Hirakawa, E., Higuchi, H. & Toyoshima, Y. Y. Processive movement of single 22S dynein molecules occurs only at low ATP concentrations. *Proc. Natl Acad. Sci. USA* **97**, 2533–2537 (2000).
- Shingyoji, C., Higuchi, H., Yoshimura, M., Katayama, E. & Yanagida, T. Dynein arms are oscillating force generators. *Nature* **393**, 711–714 (1998).
- Schnitzer, M. J. & Block, S. M. Kinesin hydrolyses one ATP per 8-nm step. *Nature* **388**, 386–390 (1997).
- Mehta, A. D. *et al.* Myosin-V is a processive actin-based motor. *Nature* **400**, 590–593 (1999).
- Neuwald, A. F., Aravind, L., Spouge, J. L. & Koonin, E. V. AAA⁺: a class of chaperone-like ATPases associated with the assembly, operation and disassembly of protein complexes. *Genome Res.* **9**, 27–43 (1999).
- Vale, R. D. AAA proteins. Lords of the ring. *J. Cell Biol.* **150**, F13–F19 (2000).
- Samso, M., Radermacher, M., Frank, J. & Koonce, M. P. Structural characterization of a dynein motor domain. *J. Mol. Biol.* **276**, 927–937 (1998).
- Silvanovich, A., Li, M., Serr, M., Mische, S. & Hays, T. S. The third P-loop domain in cytoplasmic dynein heavy chain is essential for dynein motor function and ATP-sensitive microtubule binding. *Mol. Biol. Cell* **14**, 1355–1365 (2003).
- Mocz, G. & Gibbons, I. R. Model for the motor component of dynein heavy chain based on homology to the AAA family of oligomeric ATPases. *Structure* **9**, 93–103 (2001).
- Whiteheart, S. W. *et al.* N-ethylmaleimide-sensitive fusion protein: a trimeric ATPase whose hydrolysis of ATP is required for membrane fusion. *J. Cell Biol.* **126**, 945–954 (1994).
- Bingham, J. B., King, S. J. & Schroer, T. A. Purification of dynactin and dynein from brain tissue. *Methods Enzymol.* **298**, 171–184 (1998).
- Schroer, T. A. & Sheetz, M. P. Two activators of microtubule-based vesicle transport. *J. Cell Biol.* **115**, 1309–1318 (1991).
- Sloboda, R. D. & Rosenbaum, J. L. Purification and assay of microtubule-associated proteins (MAPs). *Methods Enzymol.* **85**, 171–184 (1982).
- King, S. J. & Schroer, T. A. Dynactin increases the processivity of the cytoplasmic dynein motor. *Nature Cell Biol.* **2**, 20–24 (2000).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements R.M. acknowledges a postdoctoral fellowship from the International Human Frontier Science Program Organization. B.C.C. acknowledges support from an NIH training grant. This work was supported by a NIGMS and a CRCC grant (to S.P.G.).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.P.G. (sgross@uci.edu).

Structural basis for removal of adenine mispaired with 8-oxoguanine by MutY adenine DNA glycosylase

J. Christopher Fromme¹, Anirban Banerjee², Susan J. Huang¹ & Gregory L. Verdine^{1,2}

¹Department of Molecular and Cellular Biology, and ²Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138, USA

The genomes of aerobic organisms suffer chronic oxidation of guanine to the genotoxic product 8-oxoguanine (oxoG)¹. Replicative DNA polymerases misread oxoG residues and insert adenine instead of cytosine opposite the oxidized base. Both bases in the resulting A·oxoG mispair are mutagenic lesions, and both must undergo base-specific replacement to restore the original C·G pair. Doing so represents a formidable challenge to the DNA repair machinery, because adenine makes up roughly 25% of the bases in most genomes. The evolutionarily conserved enzyme adenine DNA glycosylase (called MutY in bacteria and hMYH in humans) initiates repair of A·oxoG to C·G by removing the inappropriately paired adenine base from the DNA backbone. A central issue concerning MutY function is the mechanism by which A·oxoG mispairs are targeted among the vast excess of A·T pairs. Here we report the use of disulphide crosslinking² to obtain high-resolution crystal structures of MutY–DNA lesion-recognition complexes. These structures reveal the basis for recognizing both lesions in the A·oxoG pair and for catalysing removal of the adenine base.

MutY (Fig. 1) belongs to a structural superfamily of proteins responsible for the base excision and repair of various genotoxic lesions. These proteins have a catalytic domain containing a

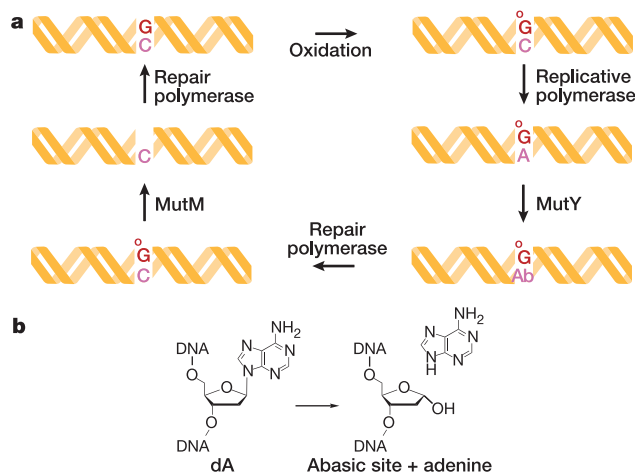


Figure 1 Pathway for oxidation and repair of guanine in DNA. **a**, The initial oxidation product, oxoG-C, is misreplicated to produce oxoG-A. MutY removes the adenine nucleobase from oxoG-A pairs, generating an abasic site product (Ab refers to the product shown in **b**). Further processing by other enzymes results in restoration of the oxoG-C pair. Repair of oxoG-C by oxoG DNA glycosylase (MutM in bacteria, hOGG1 in eukaryotes) effects removal of the oxoG nucleoside, facilitating the final phase of repair, which generates the original G-C pair. **b**, Reaction catalysed by MutY.

signature helix–hairpin–helix element, followed by a (Gly/Pro)-rich loop and a catalytically essential aspartate residue (called an HhH-GPD motif). MutY is distinct among the HhH-GPD superfamily in that it contains an additional carboxy-terminal domain that seems to be responsible for oxoG recognition, as the removal of this domain results in loss of discrimination between A·oxoG and A·G (refs 3, 4). High-resolution structures of the isolated catalytic⁵ and C-terminal domains⁶ of *Escherichia coli* MutY have been reported, but there are no structures of the intact protein or a MutY–DNA complex. The catalytic domain of MutY, including the structural [4Fe–4S] cluster, bears a strong overall resemblance to endonuclease III (EndoIII). It is intriguing that the C-terminal domain of MutY is highly similar to MutT^{4,6}, an enzyme that sanitizes the nucleotide precursor pool by hydrolysing oxo-dGTP to oxo-dGMP and inorganic pyrophosphate.

We have crystallized full-length MutY, relative molecular mass 41,000, derived from the thermophilic bacterium *Bacillus stearothermophilus*. On the basis of two preliminary structures (see Supplementary Information), we devised a strategy by which the stalled A·oxoG recognition complex was stabilized through inter-

molecular disulphide crosslinking² to a Asp144Asn variant of MutY (Supplementary Fig. S1); this approach yielded crystals that diffracted to 2.2 Å resolution, allowing us to determine the structure of the complex (Fig. 2a, b).

As anticipated, the mode of DNA interaction used by the MutY catalytic domain (Fig. 2c) is similar to that of AlkA⁷, hOGG1 (ref. 8), and in particular EndoIII (ref. 9). This domain interacts most extensively with the strand containing the substrate adenine residue, which runs through a deep cleft at the junction between the two structural modules that comprise the catalytic domain (the six-helix barrel module and the [4Fe–4S] cluster). The substrate adenine is completely extruded from the DNA helix and is inserted into an extrahelical pocket in the catalytic domain, a feature that is shared by all known DNA glycosylases that act on single-base lesions. As with EndoIII, the six-helix barrel module of MutY directly contacts the backbone of the complementary oxoG-containing strand. MutY bends the DNA substrate 55°, which is roughly the same as the bend induced by EndoIII, but less than that induced by AlkA (66°) or hOGG1 (70°). The bend is localized to the lesion, with normal un bent B-form DNA helices projecting from either side.

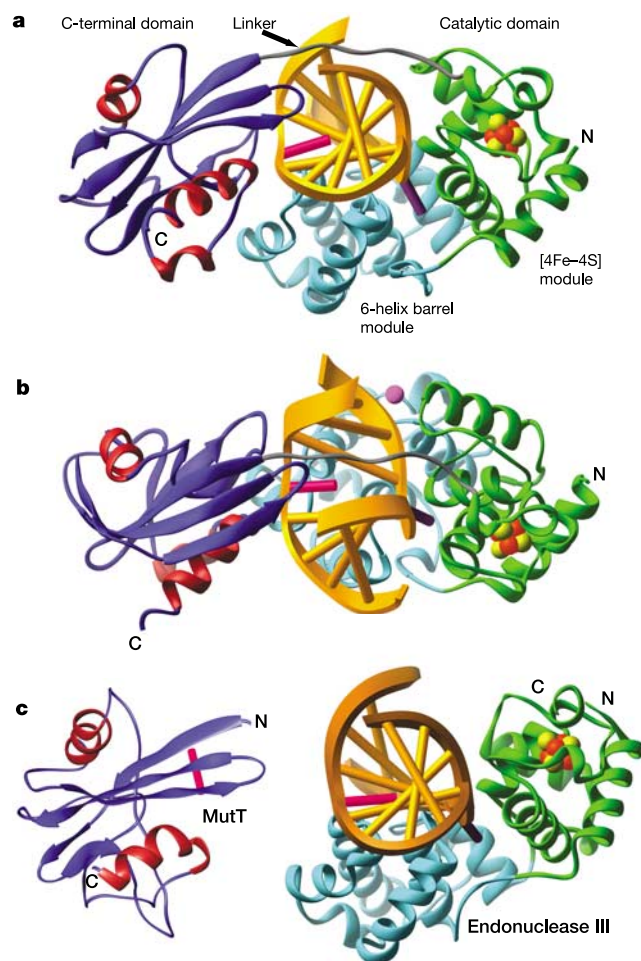


Figure 2 MutY–DNA complex. **a**, Ribbon trace of the complex. DNA is shown in gold, oxoG in magenta, the substrate adenine in purple, the [4Fe–4S] domain in green, and the six-helix barrel domain in cyan. The C-terminal domain is coloured by secondary structure (β -strands, blue; helices, red). The sulphur and iron atoms of the [4Fe–4S] cluster are yellow and orange, respectively. See Supplementary Fig. S4 for a topology diagram. **b**, Same structure as in **a**, but rotated 90° towards the reader. The purple sphere indicates a Ca^{2+} ion that is occluded in **a**. **c**, Structures of MutT¹³ and the EndoIII–DNA complex⁹. N and C termini are indicated.

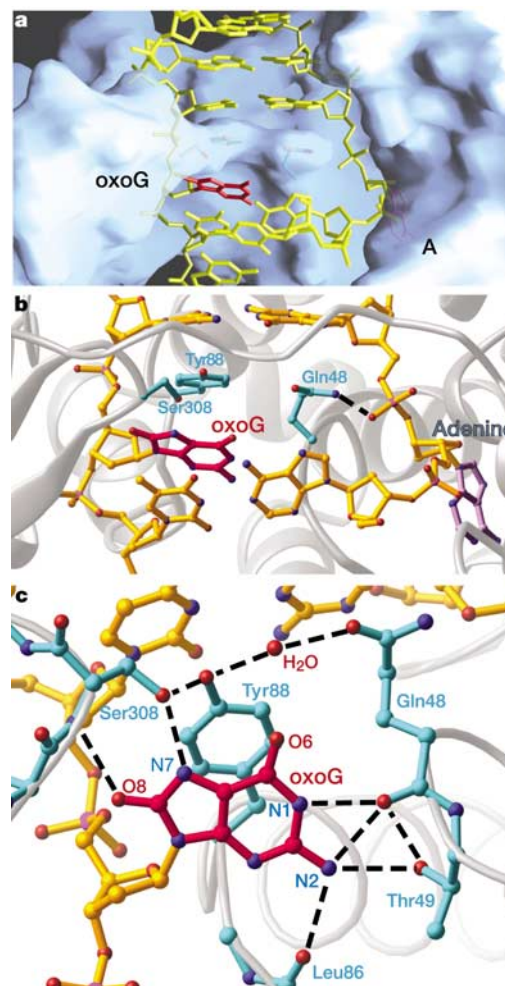


Figure 3 Enzyme–DNA interface and oxoG recognition. **a**, GRASP²⁴ molecular surface representation of the enzyme, with the bound DNA shown as a framework model. Residues Gln 48, Tyr 88 and Ser 308 are indicated and the DNA is coloured as in Fig. 1. **b**, Ball-and-stick representation of the view shown in **a**. OxoG is shown in red, the substrate adenine in purple, and the remainder of the DNA in gold. The protein backbone is presented as a grey ribbon trace, with side chains shown explicitly in cyan. **c**, Close-up view of oxoG recognition by MutY (coloured as in **a**). Broken lines indicate inferred hydrogen bonds.

Comparison of the DNA-bound (*B. stearothermophilus*, present work) and unbound (*E. coli*, ref. 5) MutY catalytic domain structures indicates that lesion recognition does not require any significant conformational changes in this domain.

The unique C-terminal domain of MutY makes extensive contacts to the oxoG-strand, making it an integral component of the overall protein–DNA interface. This domain is tethered to the catalytic domain through an extended, partially disordered linker of ten amino acids that traverses the major groove near the lesion (Fig. 2a, b). The linker is rich in lysine residues (Lys 228, Lys 230, Lys 231 and Lys 235) that are within range to make favourable electrostatic interactions with the DNA duplex, although the side chains of the middle two are disordered. The C-terminal domain not only contacts the backbone and lesion base of the oxoG-strand, but also has substantial interactions with the six-helix barrel module of the catalytic domain (contact surface area, 444 Å², Fig. 2a and Supplementary S2). These interdomain contacts, together with the linker, allow the enzyme to encircle the DNA duplex, burying 1,530 Å² of DNA surface area (including the substrate adenine).

MutY residues deeply penetrate the DNA helix, interrupting helical stacking on both strands and enforcing the sharp bend and extrahelical extrusion of the substrate adenine (Fig. 3a, b). Unlike the substrate adenine, the oxoG lesion lies completely in the DNA helix, a result that contrasts with previous conclusions derived from protein-induced changes in fluorescence intensity¹⁰. We think that

the previously observed changes reflect DNA bending rather than extrahelical extrusion of the oxoG.

Notably, the oxoG nucleoside has an *anti* glycosidic bond conformation when bound to MutY and disengaged from base-pairing, whereas its conformation is *syn* when paired with adenine and not bound to MutY^{11,12}, indicating that the oxoG must swivel roughly 180° about its glycosidic bond when bound to MutY. This action might drive extrusion of the substrate adenine from the helix, because an *anti* oxoG would experience steric clash with an intrahelical partner adenine. The aromatic ring of Tyr 88 is intercalated into the duplex on the 5' side of the oxoG, which nearly doubles the distance separating oxoG from its 5' neighbour (~6.5 Å versus ~3.7 Å). The amide-bearing side chain of Gln 48 inserts into the helical space vacated by the substrate adenine, forming π -stacking interactions directly underneath the 3'-neighbouring base and simultaneous hydrogen bonds with the 3' phosphate of the substrate adenine.

The oxoG base is recognized through an extensive array of direct contacts involving residues from the catalytic and C-terminal domains (Fig. 3c). The Watson–Crick face of oxoG is contacted by the main chain amide carbonyl of Gln 48 and the side chain hydroxyl of Thr 49, which are also hydrogen-bonded to each other. The minor groove face of oxoG contacts the main chain carbonyl of Leu 86 through hydrogen bonding. The Hoogsteen face of oxoG is recognized by hydrogen bonding between N7–H and the Ser 308 hydroxyl, which in turn is oriented and stabilized through simultaneous hydrogen bonding to the Tyr 88 hydroxyl. In addition, the O8 carbonyl of oxoG seems to accept a hydrogen bond from the main chain amide proton of Ser 308, although the geometry is not optimal. Notably, the oxoG recognition mode used by the MutT-like C-terminal domain of MutY is completely different from that of MutT itself¹³: the oxoG-binding sites in the two proteins are completely non-overlapping (compare Fig. 2a and c). This observation explains why residues of MutT that are involved in oxoG recognition (Arg 78 and Asn 119)¹⁴ are not conserved in MutY.

Our structure suggests the molecular basis for MutY's preferential recognition of oxoG versus thymine as the base-pairing partner for adenine when the concentration of T·A is 10⁵- to 10⁶-fold greater than that of oxoG·A. MutY assembles an extensive array of hydrogen-bonding functionality that converges on oxoG and engages every available face of the base. This functional array has exquisite chemical complementarity to the surface of oxoG, but is altogether unsuitable for recognizing thymine. This lack of stabilizing interactions to thymine is compounded by the substantially greater stability of the T·A base pair in duplex DNA as compared with oxoG·A (3.4 kcal mol⁻¹; ref. 15).

Although MutY can remove adenine from A·G mispairs, this substrate is probably of greater historical and mechanistic significance than of physiological relevance. The enzyme shows a greater than sixfold kinetic preference for A·oxoG over A·G (ref. 16), despite the fact that helix disruption with A·oxoG is disfavoured over A·G by 3.4 kcal mol⁻¹. Ser 308, a strictly conserved residue (Supplementary Fig. S3), seems to be primarily responsible for discriminating between oxoG and guanine. The hydrogen-bonding contact of the Ser 308 amide N–H to O8 of oxoG would be lost with guanine, and the side chain contact to N7–H would have to reverse polarity to interact with the N7 lone pair on guanine. Polarity reversal can ordinarily be readily accomplished with an isolated donor–acceptor pair, but the Ser 308 side chain participates in an intricate hydrogen-bond network with Tyr 88 and interacts through a bridging ordered water molecule with Gln 48; this whole network may be destabilized by polarity reversal at the Ser 308 side chain. Other more subtle factors, such as differences in the strength of the Tyr 88 stacking interaction, may be involved in discriminating guanine from oxoG. The calculated local dipole on the Hoogsteen face of the five-membered ring has an opposite orientation in guanine and oxoG (W. Yang and M. Karplus, personal communi-

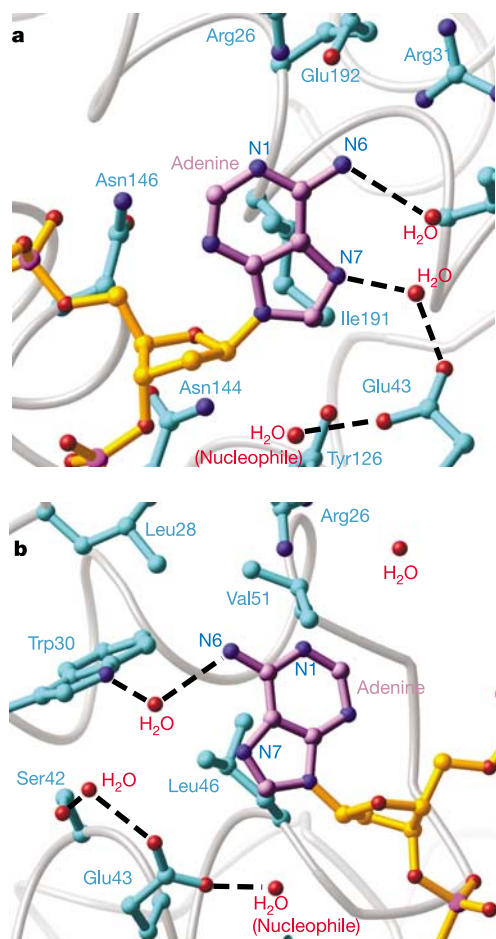


Figure 4 Close-up views of the MutY active site. **a**, View in the active-site region of the stalled oxoG·A recognition complex, with the same colouring scheme as in Fig. 3. The putative nucleophilic water is labelled. Note the lack of direct interactions with the adenine base. **b**, Same structure as in **a**, but viewed from the opposite side of the substrate base.

cation) and hence would differ markedly in the character of interaction with local dipoles on the protein.

Although the stalled recognition complex is enzymatically inactive, the structure has features that suggest the catalytic strategy used by MutY. Biochemical and computational evidence indicate that DNA glycosylases operate by a dissociative (nucleophilic S_N1 -like substitution) reaction mechanism^{17,18}, in which glycosidic bond cleavage and nucleophilic attack by water occur in discrete reaction steps. Such a mechanism can be catalysed effectively by stabilizing the oxocarbenium ion intermediate and base leaving group. Asn 144 is positioned near the O4' atom of the substrate sugar moiety, such that the corresponding, catalytically essential aspartate residue could stabilize the positively charged oxocarbenium intermediate in the base-excision step (Fig. 4); the helix-capping interaction made by Asn 144 is virtually identical to that observed for both the aspartate and asparagine variants of other superfamily members.

Two ordered water molecules are bound in catalytically relevant positions in the active site. One of these is positioned underneath the substrate sugar nearest C1' and thus is an obvious candidate for the nucleophilic water molecule that intercepts the oxocarbenium ion intermediate to form the final product. The other is hydrogen-bonded to the adenine N7, an interaction that would stabilize the base leaving group in the base-excision step. Both of these apparently catalytic water molecules lie within hydrogen-bond contact of the Glu 43 side chain, which could function as a general acid-base catalyst. This arrangement of contact partners with respect to each other and the substrate suggests that Glu 43 is initially protonated and transfers its proton to N7 through the bound water intermediate, thereby lowering the transition state energy for base excision. The resulting Glu 43 carboxylate is then primed to activate the catalytic water in the second reaction step. Consistent with this scheme, replacing the adenine N7 with C7-H yields a substrate analogue that binds to MutY but cannot be cleaved¹⁹, and mutation of Glu 43 to serine in the *E. coli* orthologue abrogates activity⁵. The exquisite arrangement of substrate atoms and catalytic functionality is only possible with a *syn*-configured adenine base.

Two polymorphisms in the *hMYH1* gene encoding the human orthologue of MutY, Tyr165Cys and Gly382Asp, are associated with multiple colorectal adenomas^{20–22}. The corresponding mutations in *E. coli* MutY result in reduced catalytic activity and relaxed discrimination for an oxoG-A substrate²³. Furthermore, the *E. coli* equivalent of MutY Tyr165Cys shows faster product release than does the wild-type enzyme. We previously highlighted the important role of the residue corresponding to Tyr 165 in the structurally related enzyme EndoIII (ref. 9) and can now comment in detail on the effects that both mutations would have on the interaction of MutY with DNA.

Tyr 165 and Gly 382 in hMYH correspond to Tyr 88 and Gly 260 in *B. stearothermophilus* MutY, both residues of which have roles in recognizing the oxoG-strand. As discussed above, Tyr 88 directly intercalates into the DNA duplex between oxoG and the nucleoside 5' to oxoG (Fig. 2). Gly 260 initiates a type-II turn in the C-terminal domain, and backbone amide nitrogen atoms of this turn form hydrogen bonds with the two phosphates immediately 5' to the oxoG nucleoside (Supplementary Fig. S2). Mutation to an aspartate at this position is likely to disrupt the structure of this turn and thus interfere with the ability of MutY to recognize faithfully oxoG in DNA. □

Methods

MutY preparation

We identified the *B. stearothermophilus* *mutY* gene by a BLAST search of the genome (<http://www.genome.ou.edu/bstearo.html>). The gene was cloned into the pET28 expression vector (Novagen), resulting in an open reading frame with an amino-terminal histidine tag and a thrombin protease cleavage site. The Asp144Asn mutation was generated by the Quikchange mutagenesis kit (Stratagene), and the Pro164Cys mutation was generated by megaprimer mutagenesis. Our expression constructs have two additional mutations, Phe347Ser and Lys357Glu, which enhance expression but retain

enzyme activity. Protein was overexpressed in BL21(DE3)pLysS cells (Novagen). Cells were lysed by sonication after the addition of 1 mM phenyl methylsulphonyl fluoride. Soluble lysates were subjected to chromatography using Ni²⁺-NTA (Qiagen), followed by MonoQ (Pharmacia). We carried out a thrombin digest to remove the histidine tag and further purified the protein on a Superdex 75 gel-filtration column (Pharmacia) in buffer comprising 20 mM Tris (pH 7.4), 200 mM sodium chloride and 5 mM β -mercaptoethanol (β -ME). Fractions containing MutY were pooled and concentrated to 15 mg ml⁻¹ (350 μ M) for storage at 4 °C.

DNA, complex formation and disulphide trapping

Using a 392 DNA synthesizer (ABI) and standard reagents, we synthesized the DNA oligomers 5'-TGTCCAXGTCT-3', where X is a reduced abasic site (phosphoramidite synthesized in our laboratory), and 5'-AAGACYTGGAC-3', where Y is oxoG (phosphoramidite purchased from Glen Research) and the italic font indicates where a modified nucleoside (from Glen Research), plus subsequent modification, was introduced for disulphide crosslinking (see Supplementary Fig. S1 for nucleoside structure and crosslinking strategy). DNA was purified by urea-PAGE, dissolved in 10 mM Tris (pH 8.0) and annealed at a concentration of 700 μ M. The DNA-protein complex was formed by mixing MutY with a 1.5-M excess of DNA duplex so that the final concentration of MutY was 175 μ M in buffer comprising 15 mM Tris (pH 7.6), 100 mM NaCl and 5 mM β -ME (only 50 μ M β -ME was used with the disulphide crosslinked complex to avoid reduction of the crosslink, and the crosslinked complex was further purified by MonoQ chromatography).

Crystallization

The complexes were crystallized by the hanging-drop method at room temperature in 100 mM Tris (pH 8.5), 14% (w/v) PEG 4000, 500 mM calcium acetate and 5 mM β -ME. Crystals appeared after several days and were allowed to grow for several weeks. Crystals were transferred to a cryoprotectant solution of 100 mM Tris (pH 8.5), 14% (w/v) PEG 4000, 500 mM calcium acetate, 5 mM β -ME and 25% glycerol, and were frozen in liquid nitrogen for X-ray data collection. To obtain the adenine-soaked complex, we soaked the crystal used for collecting data on the abasic complex in a solution of 100 mM Tris (pH 8.5), 14% (w/v) PEG 4000, 500 mM calcium acetate, 5 mM β -ME, 25% glycerol and 2.5 mM adenine at room temperature for 2 h before refreezing it for data collection.

Received 16 August; accepted 19 December 2003; doi:10.1038/nature02306.

- Lindahl, T. Instability and decay of the primary structure of DNA. *Nature* **362**, 709–715 (1993).
- Norman, D. P. G. & Verdine, G. L. Covalent trapping of protein-DNA complexes. *Annu. Rev. Biochem.* **72**, 337–366 (2003).
- Gogos, A., Cillo, J., Clarke, N. D. & Lu, A. L. Specific recognition of A/G and A/7,8-dihydro-8-oxoguanine (8-oxoG) mismatches by *Escherichia coli* MutY: removal of the terminal domain preferentially affects A/8-oxoG recognition. *Biochemistry* **35**, 16665–16671 (1996).
- Noll, D. M., Gogos, A., Granek, J. A. & Clarke, N. D. The C-terminal domain of the adenine-DNA glycosylase MutY confers specificity for 8-oxoguanine-adenine mispairs and may have evolved from MutT, an 8-oxo-dGTPase. *Biochemistry* **38**, 6374–6379 (1999).
- Guan, Y. et al. MutY catalytic core, mutant and bound adenine structures define specificity for DNA repair enzyme superfamily. *Nature Struct. Biol.* **5**, 1058–1064 (1998).
- Volk, D. E. et al. Structural similarities between MutT and the C-terminal domain of MutY. *Biochemistry* **39**, 7331–7336 (2000).
- Hollis, T., Ichikawa, Y. & Ellenberger, T. DNA bending and a flip-out mechanism for base excision by the helix-hairpin-helix DNA glycosylase, *Escherichia coli* AlkA. *EMBO J.* **19**, 758–766 (2000).
- Bruner, S. D., Norman, D. P. & Verdine, G. L. Structural basis for recognition and repair of the endogenous mutagen 8-oxoguanine in DNA. *Nature* **403**, 859–866 (2000).
- Fromme, J. C. & Verdine, G. L. Structure of a trapped endonuclease III-DNA covalent intermediate. *EMBO J.* **22**, 3461–3471 (2003).
- Bernards, A. S., Miller, J. K., Bao, K. K. & Wong, I. Flipping duplex DNA inside out: a double base-flipping reaction mechanism by *Escherichia coli* MutY adenine glycosylase. *J. Biol. Chem.* **277**, 20960–20964 (2002).
- Kouchakdjian, M. et al. NMR structural studies of the ionizing radiation adduct 7-hydro-8-oxodeoxyguanosine (8-oxo-7H-dG) opposite deoxyadenosine in a DNA duplex. 8-Oxo-7H-dG(*syn*)-dA(*anti*) alignment at lesion site. *Biochemistry* **30**, 1403–1412 (1991).
- McAuley-Hecht, K. E. et al. Crystal structure of a DNA duplex containing 8-hydroxydeoxyguanine-adenine base pairs. *Biochemistry* **33**, 10266–10270 (1994).
- Lin, J., Abeygunawardana, C., Frick, D. N., Bessman, M. J. & Mildvan, A. S. Solution structure of the quaternary MutT-M²⁺-AMPCPP-M²⁺ complex and mechanism of its pyrophosphohydrolase action. *Biochemistry* **36**, 1199–1211 (1997).
- Massiah, M. A., Saraswat, V., Azurmendi, H. F. & Mildvan, A. S. Solution structure and NH exchange studies of the MutT pyrophosphohydrolase complexed with Mg²⁺ and 8-oxo-dGMP, a tightly bound product. *Biochemistry* **42**, 10140–10154 (2003).
- Plum, G. E., Grollman, A. P., Johnson, F. & Breslau, K. J. Influence of the oxidatively damaged adduct 8-oxodeoxyguanosine on the conformation, energetics, and thermodynamic stability of a DNA duplex. *Biochemistry* **34**, 16148–16160 (1995).
- Porello, S. L., Leyes, A. E. & David, S. S. Single-turnover and pre-steady-state kinetics of the reaction of the adenine glycosylase MutY with mismatch-containing DNA substrates. *Biochemistry* **37**, 14756–14764 (1998).
- Werner, R. M. & Stivers, J. T. Kinetic isotope effect studies of the reaction catalyzed by uracil DNA glycosylase: evidence for an oxocarbenium ion-uracil anion intermediate. *Biochemistry* **39**, 14054–14064 (2000).
- Dinner, A. R., Blackburn, G. M. & Karplus, M. Uracil-DNA glycosylase acts by substrate autocatalysis. *Nature* **413**, 752–755 (2001).
- Porello, S. L., Williams, S. D., Kuhn, H., Michaels, M. L. & David, S. S. Specific recognition of substrate analogs by the DNA mismatch repair enzyme MutY. *J. Am. Chem. Soc.* **118**, 10684–10692 (1996).
- Al-Tassan, N. et al. Inherited variants of MYH associated with somatic G:C→T:A mutations in colorectal tumors. *Nature Genet.* **30**, 227–232 (2002).

21. Jones, S. *et al.* Biallelic germline mutations in MYH predispose to multiple colorectal adenoma and somatic G:C→T:A mutations. *Hum. Mol. Genet.* **11**, 2961–2967 (2002).
22. Sieber, O. M. *et al.* Multiple colorectal adenomas, classic adenomatous polyposis, and germ-line mutations in MYH. *N. Engl. J. Med.* **348**, 791–799 (2003).
23. Chmiel, N. H., Livingston, A. L. & David, S. S. Insight into the functional consequences of inherited variants of the hMYH adenine glycosylase associated with colorectal cancer: complementation assays with hMYH variants and pre-steady-state kinetics of the corresponding mutated *E. coli* enzymes. *J. Mol. Biol.* **327**, 431–443 (2003).
24. Nicholls, A., Sharp, K. A. & Honig, B. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins Struct. Funct. Genet.* **11**, 281–296 (1991).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank H. Nash for the reduced abasic phosphoramidite; Y. Korkhin for assistance with data collection and processing; M. Becker for beamline assistance; S. Bruner and J. J. Miranda for critically reading the manuscript; and Enanta Pharmaceuticals for use of their X-ray generator and detector. Some data for this study were measured at beamline X25 of the National Synchrotron Light Source; financial support for this beamline comes from the NIH and the United States Department of Energy.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to G.L.V. (verdine@chemistry.harvard.edu). Atomic coordinates have been deposited in the Protein Data Bank under accession numbers 1RRQ, 1RRS and 1RRT.

Contacts

Publisher: Ben Crowe
Editor: Paul Smaglik
Marketing Manager: David Bowen

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Naturejobs Sales Director:
Nevin Bayourni (4978)

UK/ RoW/ Ireland:

Matt Powell (4953)
Andy Douglas (4975)
Frank Phelan (4944)

Scandinavia/ Spain/ Portugal:

Evelina Rubio Håkansson (4973)

Natureevents: Silke Opstrup (4994)

France/ Switzerland:

Amelie Pequignot (4974)

Production Manager: Billie Franklin

To send materials use London
address above.
Tel +44 (0) 20 7843 4814
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Ben Lund

European Satellite Office

Germany/ Austria/ Netherlands/

Italy/ Belgium:

Patrick Phelan, Odo Wulffen
Tel + 49 89 54 90 57 11/-2
Fax + 49 89 54 90 57 20
e-mail: p.phelan@nature.com
o.wulffen@nature.com

US Head Office, New York

345 Park Avenue South,
10th Floor, New York, NY 10010-1707
Tel +1 800 989 7718
Fax +1 800 989 7103
e-mail: naturejobs@natureny.com

US Sales Manager: Peter Bless

US Advertising Coordinator:

Linda Adam

Japan Head Office, Tokyo

MG Ichigaya Building (5F),
19-1 Haraikatomachi,
Shinjuku-ku,
Tokyo 162-0841
Tel +81 3 3267 8751
Fax +81 3 3267 8746

Asia-Pacific Sales Director:

Rinoko Asami
e-mail: rasami@naturejpn.com

naturejobs

Patching a leaky pipeline

The diagnosis is clear: women are underrepresented in science. And a 'leaky pipeline' means that for those women who do move into this sector, disproportionately fewer of them will rise to the top. Two broad treatments now exist — tackle the symptoms or grapple with the cause. Dealing with the symptoms seems to be easier, but can be less effective. For example, under its Sixth Framework Programme, the European Commission is encouraging academia and industry to hire and promote more women (see *Nature* 426, 210–211; 2003). But systems of quotas can cause resentment.

Instead, treating the individual causes — such as women who temporarily step off the tenure ladder to start a family and find it difficult to climb back on — could be a more successful course of action. In the United States, organizations such as the Association for Women in Science (www.awis.org) and philanthropies including the Alfred P. Sloan Foundation (www.sloan.org) have been aggressive in attacking this problem through creative grant schemes.

More recently, European organizations have joined the fray. The European Molecular Biology Organization offers a 'restart' grant, Britain has the Dorothy Hodgkin Fellowships and the Swiss National Science Foundation awards the Marie Heim-Vögtlin grants. These programmes are beginning to yield success stories (see *Nature Med.* 10, 114–115; 2004).

But they could still be more widespread — encompassing a greater number of countries. And those programmes that do exist in individual countries might be better promoted under a 'European women in science' umbrella. Finally, 'restart' grants, as beneficial as they are, address only one cause of women's underrepresentation in science. Other issues — covert and overt, unintentional or not — still exist in institutions in many parts of the world. Addressing them will help more women to enter science, and rise to the top.

Paul Smaglik
Naturejobs editor



Contents

REGIONS

San Francisco's
biotech mission p658

CAREER VIEW

Bricks & Mortar
The Partnership for Structural
Biology
Graduate Journal
Seeking perspective
Movers
William Chia p660

WWW.NATUREJOBS.COM

Career centre
Information on the
scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

ANNOUNCEMENTS

EVENTS

San Francisco Baywatch

Biotech has a brilliant and distinguished history in the Bay Area, but the fortunes of the whole region have turned sour in recent years with the dot-com bust. The hope now is that, with some massive investment in 'interdisciplinary infrastructure' that will unite widely disparate fields, the glory years will return.

The biotech industry was born in the Bay Area in 1972 with the invention of recombinant DNA by Stanley Cohen from Stanford University and Herbert Boyer of the University of California, San Francisco (see *Nature* **421**, 456–457; 2003). The reverberations echoed around the world and helped to launch more than 600 companies and create thousands of biotech jobs in the Bay Area.

For the biotech industry to progress it needs massive investment in interdisciplinary science to bring together biologists, mathematicians, physicists, chemists, computer scientists and engineers in the same setting. If successful, this could lead to advances in diagnostics, tissue engineering and nanotechnology, which in turn will spawn new companies and jobs for the recession-hit area.

The effects of the stock-market bubble bursting in 2001 are still being felt. Venture-capital investment is down and the state budget is tight. And despite the downturn, house prices still linger in the stratosphere, making newcomers wary. Even with higher unemployment, traffic still gets snarled on the bridges connecting Bay Area campuses and companies.

But two sets of investment in academic infrastructure could help to break the economic gridlock. The first will unite 110 principal investigators from three University of California (UC) campuses — San Francisco (SF), Santa Cruz and Berkeley — under the banner of the California Institute for Quantitative Biomedical Research (QB3), on a new campus on the only piece of undeveloped real estate in San Francisco. In the second, Stanford University has brought together 40 professors for the multidisciplinary Bio-X initiative.

Barriers are being broken down, not just between disciplines but also between campuses, and between the UC system and the private Stanford University powerhouse. In turn, these developments have brought academia and industry even closer together.

The epicentre for change lies just south of the city. Mission Bay used to be dominated by warehouses and railroad switchyards — the kind of place that even the Beat poets who helped make the city famous in the



Bridging the gap: researchers at Genentech Hall, part of the University of California, San Francisco, are encouraged to mingle with colleagues from other disciplines.

1950s might have shied away from. Now it's a massive construction site. The city of San Francisco gave UC 45 of the area's 300 or so barren acres, partly because of fears that UCSF might move its operations elsewhere, and partly to spur retail, commercial and residential development.

So far, the experiment has been a qualified success. Genentech Hall, which houses 30 research groups and some 500 chemists, biologists and structural biologists, opened last May. A developmental-biology and human-genetics building for another 500 scientists opens next month. And the QB3 building, which will house instrumentation and a biomedical-engineering programme, will open next year. If all goes according to plan, the site will eventually host 9,100 faculty members, graduate students, postdocs and support staff in 20 buildings.

CREATING TRANSPARENCY

Three other buildings — for brain research and systems neuroscience, cardiovascular diseases and cancer — are on the drawing board. But problems with the state budget have put their construction schedules under threat. And short-term financial problems have ignited some concerns about overheads and operating costs (see *Nature* **427**, 274; 2004).



But the projects aren't just about putting up new buildings. "Mission Bay offers the opportunity to reshape a whole institution," says Keith Yamamoto, professor of biochemistry at UCSF and executive vice-dean of the university's medical school. At Genentech Hall, administrators deliberately organized labs into 'neighbourhoods' where people from different disciplines can mingle and departments can support one another (see *Nature* **424**, 718–720; 2003).

The biggest challenge, according to Yamamoto, is to avoid overlap and unnecessary fiefdoms, as the three campuses of QB3 share four broad research areas: bioengineering and biotechnology; bioinformatics and computational biology; structural and chemical biology; and experimental genomics, proteomics and biochemistry. "Some integration is happening," Yamamoto says. "And any sort of integration will be seen as a step in the right direction."

Perhaps one of the best examples is the bioengineering graduate programme, supported jointly by UC Berkeley and UCSF. Over four years it has almost tripled in size to 136 students. And faculty appointments are rising just as fast, says Sarah Nelson, UCSF's graduate adviser on the programme.

One sign of this growth is apparent at UC Berkeley's campus, a 20-minute drive from Mission Bay across the Bay Bridge, where there is a giant hole. In 2006, it will be filled by Stanley Hall, a \$168-million building dedicated to bioengineering and nanotechnology that will be home to 40 principal investigators and 13 high-resolution nuclear magnetic resonance machines.

Back across the bridge and 40 minutes south of Mission Bay, the James H. Clarke Center at Stanford, a \$140-million 'glass fishbowl', gives a new meaning to the expression 'transparency in science'. The building's

glass walls will make it easy for scientists to see what's going on in other labs. It opened in October and is designed, like the QB3 buildings, to foster interaction.

Russ Altman, associate professor of genetics and medicine at Stanford, says that the social engineering has worked out better than he anticipated. Altman leads the Pharmacogenetics and Pharmacogenomics Knowledge Base, a database that helps scientists to understand how drugs react in people with different genetic make-ups. He found it took only six months to establish a multidisciplinary team and complete a 400-page biocomputing proposal.

Nobel laureate Steve Chu, a biophysicist at Stanford and a Bio-X scientist, says that the interactions have been organic, dictated by the direction the science takes. After pioneering 'optical tweezers', Chu decided to use physical tools to play with biological molecules, rather than be limited to one discipline. Now he uses structural biology and imaging to investigate how small genetic mutations can make a protein fold into different shapes.

ON THE MOVE

Such interactions are being seen from beyond the glass walls of the Clarke Center. "People in other departments are much more aware of what's going on, instead of having isolated pockets of people doing their own thing," says Chu. And the university is now asking Bio-X affiliates to help recruit people into other departments, in order to build up complementary skills and interests throughout the university.

That spirit of collaboration is evident 40 minutes south by car at UC Santa Cruz. The campus offers QB3 expertise in computational biology and engineering. David Haussler, a Santa Cruz bioinformatics professor who played a key role in the Human Genome Project, heads the QB3 there. But despite the improving infrastructure — one research lab went up last year and an engineering building is scheduled for completion at the end of the year — Haussler often has to visit other campuses.

He isn't alone. Chu travels from Stanford to Mission Bay every month for an 'RNA club' whose dozen or so members hail from Stanford, UCSF, Berkeley and Santa Cruz. And Haussler, adviser to the pharmacogenetics database project, often drives to Stanford to meet Altman. Altman, in turn, frequently goes to UCSF to meet with Kathleen Giacomini, chair of its biopharmaceutical-sciences department.

The mass transit extends to industry. Ellen Filvaroff, a scientist at Genentech in South San Francisco, visits university campuses and brings undergraduates, graduate students and postdocs back to Genentech to give them a realistic idea of what happens in industry. She thinks that Mission Bay's location, so close to South San Francisco's concentration of biotech hubs, is a boon.

Charles Craik, QB3 director of chemistry and chemical biology, runs a similar course at Genentech Hall. It draws students from all the campuses in the area, as well as from biotech companies, and teaches them the basics of the business, such as venture capital, business plans and intellectual property.

Once the local economy has regained its feet, such interactions between the universities and industry could help the Bay Area make history again.

Paul Smaglik is editor of *Naturejobs*.

GRADUATE JOURNAL

Seeking perspective

Most researchers get their first career tips from their supervisors. But that can be limiting — supervisors tend to skew their advice towards their own research area. Senior colleagues at meetings and conferences are another source of advice, but this relies on your networking skills and depends on who you meet.

A more formalized approach to mentoring can be beneficial. I remember a session with a professional adviser I had after I finished my undergraduate degree. He analysed my interests and abilities, and then recommended a few career paths likely to fit my personality. Unfortunately, such sessions tend not to be available at European universities. Some students are changing that, through regional clubs or international student organizations, such as the Young European Biotech Network.

With some effort, these organizations could help to promote mentoring offices within universities or organize occasional career panels with scientists from various areas. Tapping into university alumni working in industry to advise graduate students on opportunities could save work. Making those experiences available from a central source would result in a broader spectrum of subjects and therefore provide a better perspective on career options. ■

Philipp Angerer is a second-year PhD student in biotechnology at the Swiss Federal Institute of Technology (ETH) in Zurich, Switzerland.

BRICKS & MORTAR

The Partnership for Structural Biology

Structural biology in Europe is getting a boost from a new partnership of international laboratories all based in Grenoble, France.

The European Synchrotron Radiation Facility (ESRF), the European Molecular Biology Laboratory (EMBL), the Laue-Langevin Institute (ILL) and the Institute of Structural Biology (IBS) are combining their expertise and instruments to create a single centre for structural proteomics — the Partnership for Structural Biology (PSB). A 3,600-m² building, which should be built by the end of 2005, will house integrated PSB teams as well as a new institute for structural virology financed by the Joseph Fourier University in Grenoble.

The partnership brings together EMBL's expertise



Partnered for proteomics: the synchrotron light source at Grenoble.

in molecular biology with the investigatory powers of the ESRF's third-generation synchrotron radiation source, the ILL's high-intensity neutron flux capacity and the facilities at the IBS for nuclear magnetic resonance and electron microscopy.

"Today's high-throughput structural biology requires robotics and new instrumentation. No single organization has all the necessary expertise or technology," says Stephen Cusack of EMBL and current chairman of the PSB.

The ESRF is contributing a highly automated macromolecular

crystallography beamline to the partnership. "Researchers will have more time to concentrate on the science if they can be freed from tedious jobs through automation," says Sine Larsen, life-sciences director at the synchrotron.

Through the PSB, all of the participating institutes are increasing their commitment to structural biology to meet modern challenges in the field. For example, the ILL plans to increase its life-sciences research from the current 5–10% to 25% of its activity. ■

Sally Goodman is a freelance science writer based in Paris.

► <http://psb.esrf.fr>

MOVERS

William Chia, director, Temasek Life Sciences Laboratory, Singapore



When developmental biologist William Chia returns to Singapore this May, he is hopeful that his previous experience in the city-state will help him to distance a resurrected institute from its fractious past.

Temasek Life Sciences Laboratory (TLL) arose from the wreckage of the Institute of Molecular Agrobiolgy (IMA) in August 2002. Once located on the

campus of the National University of Singapore, the IMA was forced to close in September 2001 amid criticism that the country — which imports most of its food — did not need agricultural research and should instead concentrate on biomedicine. Advocates of the IMA in the research community argued that the fundamental research covered more than just applied agricultural research, and many outside observers saw the episode as a clash of egos among some of Singapore's scientific élite. The IMA's 24 principal investigators were caught in the middle. In the end, half of them merged with the Institute of Molecular and Cell Biology and the other half came together to form the TLL.

Chia hopes that his combination of experience with Singapore's culture and his international perspective, gained from working in London, will allow him to help the TLL to start anew. His goal is to shape a research community that is

open and keen to collaborate with researchers within Singapore and elsewhere. "I wanted to do something worthwhile for like-minded scientists," says Chia.

Chia says that the TLL has broad-ranging fields of research: "We will focus on cell biology, development and the interface between the two." To do this, TLL researchers will use "model systems across the board", Chia says. He is clear about how to avoid the ambiguity that plagued the IMA. "We will have plant research, but it is certainly not an agrobiotechnology institute," he explains.

The TLL is still recruiting, as Chia plans to expand it from the current 16 groups and 148 researchers up to 25 groups and about 200 researchers. With all this ahead of him, Chia still hopes to continue his research into the development of the nervous system. "It's a tremendous opportunity," he says. "All the resources are here." ■

CV

2001–2004: Wellcome Trust principal research fellow and professor of developmental genetics at the Medical Research Council's Centre for

Developmental Neurobiology, King's College London

1998–2001: Deputy director, research and academic matters, Institute of Molecular and Cell Biology, Singapore

1995–2001: Professor of biochemistry, National University of Singapore

1990–1994: Associate professor, Institute of Molecular and Cell Biology, Singapore